Leveraging Artificial Neural Networks for Modeling Hydrogeological Time Series

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Andreas Wunsch

aus Mühlacker

Tag der mündlichen Prüfung:
19. Juli 2022

Referent: Prof. Dr. Nico Goldscheider
Korreferentin: Prof. Dr. Anne Johannet

Karlsruhe 2022

Leveraging Artificial Neural Networks for Modeling Hydrogeological Time Series

Doctoral Thesis

by

Andreas Wunsch

Karlsruhe, February 2022

Abstract

In solving global water challenges, such as the sustainable management and use of available groundwater resources, finding new, efficient, and easily transferable modeling approaches is crucial. Machine learning methods such as artificial neural networks (ANNs) are particularly suitable for this purpose. They can autonomously learn and leverage relevant relationships from larger data sets of suitable variables. After achieving major successes in several other fields, ANNs and especially their subset of deep learning models are becoming more and more successful in the hydrological sciences. This thesis investigates the use of ANNs to model and predict hydrogeological time series. Four studies constitute the main part of this work and demonstrate how ANNs can contribute to solving different problems in this research domain.

Groundwater hydrograph clustering is useful for identifying spatial and temporal dynamic patterns, which helps to characterize aquifer systems, identify influencing factors, and develop effective groundwater management strategies. Therefore, in the first study, an unsupervised clustering approach based on self-organizing maps is developed, capable of effectively grouping heterogeneous hydrograph datasets based on time series dynamics. A feature-based approach helps to robustly characterize hydrograph dynamics with variable data quality (e.g., data gaps or different periods). Using a data set of about 1800 weekly groundwater hydrographs, the application of the developed approach is successfully demonstrated in the Upper Rhine Graben area in Germany and France. Results show that groundwater dynamics are influenced by a variety of factors that superimpose spatially and temporarily, and often are hard to separate. Nevertheless, some clusters are clearly connected to specific external controlling factors, such as intensive groundwater management in the northern part of the study area.

Next, a detailed model comparison of different ANNs for groundwater level prediction tasks follows. The study compares nonlinear autoregressive exogenous models (NARX), long short-term memory networks (LSTM), and convolutional neural networks (CNN), each for both sequence-to-value and sequence-to-sequence prediction tasks. Furthermore, the models use only a few widely available and easy-to-measure meteorological input variables, which ensures the high transferability of the approach. All models show good predictive capabilities,

however, NARX are, on average, the most precise ones, followed closely by CNNs, LSTMs are last. For practical applications, CNNs appear best overall because they are less dependent on the random network initialization than NARX and much faster to compute than both recurrent alternatives. At the same time, they achieve high performance and can be implemented flexibly.

The subject of the subsequent study is the development of groundwater levels in Germany in the context of the climate crisis. Climate data from three climate scenarios (RCP2.6, 4.5, and 8.5) form the basis to model the direct influence of climate on groundwater using a CNN-based approach. Focusing on the direct influence means the study does not consider indirect influencing factors that are highly uncertain in the future, such as anthropogenic groundwater extractions or vegetation and land-use changes. While future developments under the optimistic RCP2.6 and the intermediate RCP4.5 result in less pronounced and fewer significant trends, the pessimistic RCP8.5 causes significantly declining groundwater levels trends for most sites, revealing a spatial pattern of stronger decreases in the northern and eastern part of Germany. The positive influence of mitigated greenhouse gas emissions is evident in the results of RCP2.6. Still, groundwater levels decrease across Germany, depending on the investigated climate model.

Finally, this thesis investigates ANNs for modeling karst spring discharge. Both the existing CNN approach and a new 2D-model that allows for direct processing of spatially distributed input data are deployed. The latter can potentially overcome the problem of limited meteorological data availability in karst catchment areas. Both approaches achieve accurate modeling results in all three test areas and partly exceed the results of already existing approaches. None of the approaches is superior in terms of accuracy. However, apart from a considerably increased computation time, the data's spatially, and temporally complete nature and the associated number of available input variables are key benefits of the 2D-approach. The 2D-models learn relevant parts of the input data automatically, and a spatial input sensitivity analysis demonstrates their usefulness to localize the position of karst catchments.

Kurzfassung

Bei der Lösung globaler Herausforderungen, wie der nachhaltigen Bewirtschaftung und Nutzung der verfügbaren Grundwasserressourcen, ist die Entwicklung neuer, effizienter und leicht übertragbarer Modellierungsansätze von entscheidender Bedeutung. Hierfür bieten sich vor allem künstliche neuronale Netze (KNN) an, die als Verfahren des maschinellen Lernens selbstständig relevante Zusammenhänge aus größeren Datensätzen geeigneter Parameter lernen und nutzen können. Die vorliegende Arbeit untersucht die Nutzung von KNN zu Modellierung und Vorhersage von hydrogeologischen Zeitreihen. In vier Studien, die den Hauptteil dieser Arbeit bilden, werden verschiedene Fragestellungen entwickelt und deren Lösbarkeit mit Hilfe von KNN demonstriert.

Das Clustern von Ganglinien ist eine Möglichkeit räumliche und zeitliche Muster der Grundwasserdynamik zu erkennen. Dies ist wichtig um Aquifere zu charakterisieren, Einflussfaktoren zu identifizieren und effektive Bewirtschaftungsmethoden zu entwickeln. Aus diesen Gründen wird in der ersten Studie auf Basis von Self-Organizing Maps ein Clustering Verfahren entwickelt, mit dessen Hilfe sich in heterogenen Datensätzen von Grundwasserganglinien solche mit ähnlicher Dynamik gruppieren lassen. Das Verfahren nutzt zur Charakterisierung der Grundwasserdynamik sogenannte Features, die auch die Verarbeitung von Ganglinien mit variabler Datenqualität ermöglichen. Anhand eines Datensatzes von ca. 1800 wöchentlichen Ganglinien wird die Anwendung im Oberrheingraben in Deutschland und Frankreich erfolgreich demonstriert. Eine Analyse der Clusterergebnisse zeigt, dass sich externe Einflussfaktoren räumlich und zeitlich komplex überlagern und eine Trennung häufig nicht möglich ist. Dennoch sind einige Cluster eindeutig auf externe Faktoren (z.B. Grundwasserbewirtschaftung) zurückzuführen.

Es folgt ein detaillierter Vergleich verschiedener KNN Modelle zur Grundwasserstandsvorhersage. Untersucht werden hierbei Nonlinear Autoregressive Models with Exogenous Inputs (NARX), Long Short-Term Memory Networks (LSTM) und Convolutional Neural Networks (CNN) sowohl jeweils für Einzelwert- als auch Sequenzvorhersagen. Als Eingangsdaten werden nur wenige, aber dafür weithin verfügbare und leicht zu messende meteorologische Parameter verwendet, wodurch die breite Übertragbarkeit des Ansatzes gewährleistet ist. Es zeigt sich, dass alle Modelltypen grundsätzlich gute Prognoseeigenschaften aufweisen und

NARX hierbei in der Regel die präzisesten Vorhersagen treffen, dicht gefolgt von CNNs. Für die praktische Anwendbarkeit zeigen CNNs insgesamt das größte Potenzial, da diese eine geringere Abhängigkeit von der pseudorandomisierten Netzinitialisierung als NARX sowie eine vielfach höhere Berechnungsgeschwindigkeit aufweisen als beide rekurrenten Alternativen. Dabei erreichen CNNs dennoch eine hohe Güte und sind gleichzeitig flexibel implementierbar. CNNs bilden daher die Grundlage für weitere untersuchte Fragestellungen.

Die nachfolgende Studie untersucht die Entwicklung der Grundwasserstände in Deutschland im Kontext des Klimawandels. Hierfür werden auf Basis von CNNs und anhand von Temperatur und Niederschlag aus drei Klimaszenarien (RCP2.6, 4.5 und 8.5) die zukünftigen Grundwasserstände an 118 ausgewählten Messstellen in Deutschland modelliert und der direkte Einfluss des zukünftigen Klimas abgeschätzt. Wichtige sekundäre Faktoren wie anthropogene Einflüsse, werden jedoch nicht in die Simulationen mit einbezogen. Unter RCP8.5 (pessimistisches Szenario) sind flächenhaft und ausgeprägt fallende Grundwasserstände zu erwarten, mit einem räumlichen Muster von stärkeren Abnahmen vor allem in Nord- und Ostdeutschland. Ebenfalls abnehmende Trends zeigen die Ergebnisse für die optimistischeren Szenarien RCP2.6 und RCP4.5, jedoch mit vergleichsweise wenig signifikanten Veränderungen. Hier wird der positive Einfluss der verminderten Treibhausgasemissionen deutlich, jedoch werden auch noch für das optimistischste Szenario RCP2.6 in einigen Projektionen deutschlandweit abnehmende Grundwasserstände festgestellt.

Abschließend stehen Karstquellschüttungen im Fokus der Arbeit. Zur Modellierung werden zum einen die vorhandenen CNN Ansätze herangezogen, zum anderen wird ein ebenfalls auf CNNs basierender 2D-Ansatz entwickelt, der die direkte Verarbeitung von flächenhaften Rasterdaten als Inputs erlaubt. Hierdurch lässt sich vielfach das Problem der ungenügenden Datenverfügbarkeit von meteorologischen Eingabedaten im Einzugsgebiet lösen. Beide Ansätze zeigen in allen Testgebieten sehr gute Ergebnisse und übertreffen teils die Ergebnisse bereits existierender Modelle. Der direkte Vergleich zwischen herkömmlichem und flächenhaftem Modellierungsansatz erlaubt kein abschließendes Urteil zur Überlegenheit einer der beiden Ansätze hinsichtlich der Genauigkeit der Ergebnisse. Die räumliche und zeitliche Vollständigkeit der Eingabedaten ist jedoch ein schwerwiegender Vorteil des flächenhaften Ansatzes. Weiterhin zeigt der flächenhafte Ansatz Potenzial für die Lokalisierung und, bei entsprechender Datenverfügbarkeit und Weiterentwicklung des Ansatzes, auch für die Abgrenzung von Quelleinzugsgebieten im Karst.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

List of important abbreviations and acronyms. All are additionally introduced in the text on their first occurrence.

AI	artificial intelligence
ANN	artificial neural network
API	application programming interface
asl	above sea level
bgl	below ground level
CA	cluster analysis
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CH	Caliński-Harabasz criterion
CNN	convolutional neural networks
DF	driving force
DL	deep learning
DM	density merge
DR	density refinement
DS2L	density-based simultaneous two-level algorithm
DWD	Deutscher Wetterdienst (German meteorological service)
E	evaporation
FD	feedback delays
FFN	feedforward neural network
GHG	greenhouse gas
GP	governing parameter
GPM	global precipitation measurement
GPU	graphics processing units
GRU	gated recurrent unit
GW	groundwater
GWL	groundwater level
HC	hierarchical clustering
HP	hyperparameter

HPD	clustering feature: high pulse duration
ID	input delays
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
KGE	Kling-Gupta efficiency
KNN	Künstliche Neuronale Netze (ANN)
LPD	clustering feature: long pulse duration
LRec	clustering feature: longest recession
LSTM	long short-term memory networks
Max	Maximum
Med01	clustering feature: Median scaled to [0,1]
Min	Minimum
ML	machine learning
MLP	multilayer perceptron
MR	McClain-Rao criterion
NARX	nonlinear autoregressive exogenous models
NLP	natural language processing
nS	new snow
NSE	Nash-Sutcliffe efficiency
NTH	neighborhood threshold
P	precipitation
p1, . . . , p6	(climate) projection 1, . . . , projection 6
P52	clustering feature: annual periodicity
PCA	principal component analysis
PCHIP	piecewise cubic hermite interpolating polynomial
PET	potential evapotranspiration
PI	persistency index
Pr	process
r	Pearson correlation coefficient
R^2	coefficient of determination, here: squared Pearson r
Rad	surface shortwave downwelling radiation
RBI	Richards-Baker index
rBias	relative Bias
RCP	representative concentration pathway
ReLU	rectified linear unit
RF	random forest
rH	relative humidity
RISE	Randomized Input Sampling for Explanation
RL	(chapter I) reinforcement learning

RL	(chapter II) Ratkowsky-Lance criterion
RMSE	root mean squared error
RNN	recurrent neural network
RQ	research question
RR	range ratio
rRMSE	relative root mean squared error
$\overline{R_W}$	weighted intra-cluster correlation
S	snow
SB	clustering feature: seasonal behavior
SEM	standard error of the mean
seq2seq	sequence to sequence
seq2val	sequence to value
SF	snowfall
SHAP	SHapley Additive exPlanations
Skew	clustering feature: skewness
SL	supervised learning
SMLT	snowmelt
SOM	self-organizing map
SWMM	Storm Water Management Model
SWVL1	volumetric soil water of layer 1
SWVL2	volumetric soil water of layer 2
SWVL3	volumetric soil water of layer 3
SWVL4	volumetric soil water of layer 4
T	temperature
T_{max}	maximum temperature
T_{min}	minimum temperature
Tsin	sinusoidal curve fitted to temperature data
UL	unsupervised learning
URG	Upper Rhine Graben
W	river water level
XAI	explainable AI
Yvar	clustering feature: yearly variance

Chapter I

Introduction and Overview

1 General Motivation

Groundwater (GW) is the major source of freshwater worldwide. At least half of the global population uses groundwater for drinking water supplies (WWAP, 2015), and it constitutes a substantial amount of global irrigation water (FAO, 2010). In the future, we will have to deal with severe changes in water availability at all spatial and temporal scales, particularly due to the climate crisis and its undoubted consequences. Already, four billion people experience water scarcity for at least one month a year (Mekonnen and Hoekstra, 2016), and billions of people worldwide do not have access to safely managed drinking water, according to the United Nations (UN-ECOSOC, 2021). However, not only the climate crisis challenges water resources. Global water consumption increased more than twice as much as population growth in the 20^{th} century, and other factors such as water pollution, degraded ecosystems, and lack of policy cooperation further exacerbate the problem (UN-ECOSOC, 2021). There is an urgent need to fundamentally transform the way we manage the Earth's limited water resources to overcome these problems (UN-Water, 2020). Modeling approaches in hydrogeology are crucial in this regard, as they are necessary to develop sustainable freshwater management strategies. From models, we can derive knowledge about hydrogeological systems' quantitative or qualitative state while using only pointwise observations, such as groundwater level measurements in distinct wells. High-quality models are further essential to calculate usage scenarios and make predictions of future developments on different time scales. Representing complex real-world hydrogeological systems in physical numerical models is a common modeling strategy. However, it requires much effort due to the substantial knowledge about systems properties necessary to parameterize the model. Alternative approaches such as lumped parameter models require less effort; however, finding an adequate simplification of the system with an acceptable error still needs distinct domain knowledge. Both concepts thus lack transferability, and the benefits of finding new, efficient, and transferable modeling approaches for the development of sustainable groundwater management actions are therefore obvious.

Only little domain knowledge and system understanding is necessary to employ artificial neural networks (ANNs). They offer a promising alternative to above-mentioned approaches because they learn from data alone and establish relevant relations automatically. ANNs compose of layer-wise ordered and interconnected artificial neurons, mimicking the basic functioning of the human nervous system. They represent a branch of machine learning (ML), a term that is sometimes synonymously used for artificial intelligence (AI). However, AI is a vaguely defined and even wider field, in which ML is only one aspect. Generally described, ML is the process of pattern extraction from data (Goodfellow et al., 2016). The data representation for such pattern extraction comprises many pieces of information called features. One common approach is to apply human knowledge to design or extract such features for the successful application of ML algorithms (feature engineering). For many tasks, which are intuitive and straightforward to humans, such as speaking and understanding a language, this poses an incredibly complex challenge because the underlying knowledge is hard to describe in a formal way (Goodfellow et al., 2016). The solution is to apply representation learning (also called feature learning), which means that ML algorithms not only learn to process existing representations but also learn to automatically extract the features from the data necessary to build such representation (Bengio et al., 2014) (Figure I.1). To perform representation learning, we can use ANNs with multiple layers, known as deep learning (DL) models. We can think of depth roughly regarding the number of layers in an ANN, meaning there are shallow (very few layers) and deep (many layers) ANNs. However, there is no distinct definition of what number of layers or what depth qualifies a model as deep (e.g., Arnold and Tilton, 2019; Schmidhuber, 2015).

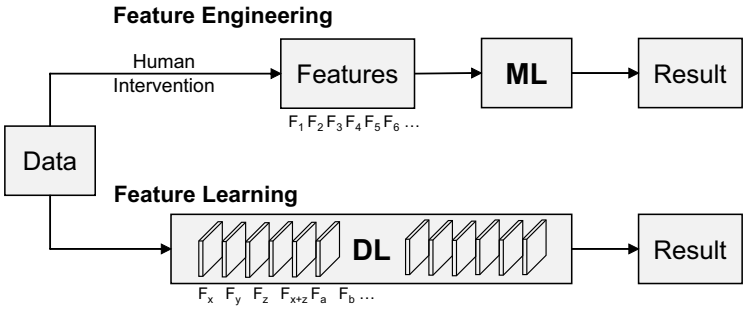


Figure I.1: Classical ML approach of feature engineering using human knowledge versus DL approach of feature learning in multiple ANN layers.

Feature learning in deep models means that complex representations are learned as a composition of more basic and abstract representations (Goodfellow et al., 2016). For example, DL models learn to recognize a human face in terms of edges, corners, contours, and specific parts (e.g., eyes) rather than in its totality. Usually, the number of layers within a network corresponds to the level of abstraction (Goodfellow et al., 2016). Based on these concepts,

DL is able to tackle real-world problems with unmatched flexibility and accuracy, which is a major reason why DL has fundamentally transformed how we live and interact with each other or machines in the last years, be it in our personal or professional lives. Prominent day-to-day examples may be natural language processing (NLP) and the resulting advances in speech recognition or intelligent product recommendations in online shopping and media streaming. Generally, it should be challenging to identify any branch of technology that ML or ANNs do not yet influence.

In recent years, the amount and variety of available data in the natural sciences have become increasingly large, which is both an opportunity and a challenge at once (Shen, 2018a). The number of sensors in our environment is growing immensely, and at the same time, they provide ever more accurate and higher-resolved data. Also, more and better spatially distributed data are available, such as from Copernicus, the European Union's earth observation program (European Union, 2022). This applies especially to meteorology, which has particular relevance for new possibilities in hydrogeological modeling by better representing major driving forces in the groundwater domain. Above all, access to data has improved considerably in recent years. Many scientists and institutions now follow the idea of open data, which increases the exchange of data and makes more and more data openly accessible. However, conventional models are often limited in their capability of incorporating large amounts of data and thus cannot fully profit from these developments. Such limited capability can either originate from a limitation in knowledge or model capability (Shen, 2018b). Furthermore, they heavily depend on human expertise for individual calibration and customization or preliminary feature extraction from data. Thus, while sticking exclusively to such approaches, we might not even be aware that we underutilize the newly available data, as we have no idea what abstract information could be additionally extracted from large data sets (Shen, 2018a; Shen et al., 2018). ANNs are generally capable of processing such large amounts of data without the need for explicit formulations of variable relations, while often even being able to generalize the gained knowledge to new instances successfully. When it comes to information extraction from raw data, representation learning in DL models can be applied, a substantial advantage compared to other ML methods (even shallow ANNs). Given these abilities, ANNs in general, and DL models in particular, are well suited to model complex real-world systems using the available data.

One drawback of ANNs is that they are usually considered black-boxes. Despite all recent developments in explainable AI (XAI), which aims for interpretability and understanding of AI models, the guideline remains that ANNs are particularly suitable as long as a mere input-output relation is of interest. There are generally two approaches to increase trust in such a black-box decision. One is to understand the model's decisions, which is the goal of XAI

methods. Understanding decisions is crucial because (i) ANNs can learn (physically) wrong or undesired relations, and (ii) we know that ANNs (in this context usually DL models) can be tricked by so-called adversarial attacks, which are smallest changes of the input data leading to erroneous results (Szegedy et al., 2014). Adversarial attacks are a known issue and active area of research, for example, in computer vision. It is worth noticing that XAI in general still faces fundamental problems as to whom a model should be explainable and how. For example, many XAI methods may be helpful for AI researchers but not for users whose expertise comes from another field (see, e.g., Gerlings et al., 2022, for examples in healthcare). Besides XAI methods, another way to increase trust is to incorporate additional knowledge to ensure the model does things for the right reasons. Usually such knowledge are constraints, e.g., laws of physics, which is also known as inductive bias (Mitchell, 1980). Generally speaking this term describes constraints that allow for prioritizing one solution over another. The major challenge here is to improve the solution search without diminishing the model performance by introducing constraints that are too strong (Battaglia et al., 2018).

ANNs were already used to model hydrogeological time series since the '90s, when for example Johannet et al. (1994) first implemented a neural network to simulate karst spring discharge. Shortly after, Maier (1995) comprehensively demonstrated the modeling of multivariate water quality time series and also provided an extensive literature review on other related ANN applications from the early '90s, such as water demand forecasting. Since then, the number of publications and applications of ML in hydrogeology increased ever stronger; however, specifically, the application of ANNs rather took part gradually, which has dramatically changed in the last years. The success of DL methods in various disciplines, the growing amount of available data, and the widely accessible computational power caused the application of ANNs to reach the mainstream of worldwide research efforts in the water-related sciences. Older (shallow) model architectures like feedforward neural networks (FFN)/multilayer perceptrons (MLP) and simpler forms of recurrent neural networks (RNN) were the commonly applied methods for many years and still have many adequate applications. However, DL approaches such as long short-term memory networks (LSTM) (Hochreiter and Schmidhuber, 1997), convolutional neural networks (CNN) (LeCun et al., 2015) and other successful model architectures from various disciplines are now far more popular and more frequently used. A quite recent literature review on applications in groundwater level (GWL) modeling is provided by Rajaee et al. (2019); Shen (2018a) and Shen et al. (2018) give a broader overview on DL and its relevance for water resources scientists. A short overview of the historical background of ANNs in general, follows in the subsequent section.

2 Historical Background and Important Concepts

2.1 ANN History

Neurosciences are still considered an important source of inspiration for some DL researchers today (Goodfellow et al., 2016) and were even more essential in the beginning of ANN history, which is generally considered to start in the 1940s. At the time, McCulloch and Pitts (1943) translated the concept of neurons from neuro- to computer sciences and showed that the so-called McCulloch-Pitts-Neuron could basically compute any logical function. However, automated learning was not yet possible until, at the end of this decade, Hebb (1949) introduced the foundation for modern learning algorithms using dynamic instead of static weights to connect neurons (Hebbian learning). In the late 1950s, the perceptron model by Rosenblatt (1958) enabled first practical pattern recognition applications, as it was able to learn to distinguish between different categories of input data. Subsequently, ANN research stagnated for more than a decade and was even sometimes considered a dead-end (Hagan et al., 2014), which marked the end of the first period of flourishing ANN research. Some difficulties, such as the lack of computational power, seemed to be overcome in the 1980s, the starting point of the second period of ANN research making great strides. Probably most important was the development of the backpropagation training algorithm by several researchers, gaining broad attention by the work of Rumelhart et al. (1986). Goodfellow et al. (2016) refer to this second period as the *connectionism* and describe its basic idea as the belief that a network of interconnected simpler computational units would be able to achieve intelligent behavior. During this time, Fukushima and Miyake (1982) developed the Neocognitron, the predecessor of convolutional neural networks, a popular DL model architecture of the present. In the early 1990s, researchers formulated major problems with time series forecasting, such as the vanishing and exploding-gradient-problems (Bengio et al., 1994; Hochreiter, 1991). This led to the development of another very popular DL architecture of today, namely long short-term memories (Hochreiter and Schmidhuber, 1997), which can overcome some of these problems. Despite all achievements, this second period of success in ANN research ended in the mid-1990s, when other ML models such as kernel machines and graphical models solved important tasks. At the same time, ANN research did not fulfill its own ambitious goals, formulated primarily by companies seeking to attract investors (Goodfellow et al., 2016). The third and still lasting period of successful AI research may have started in 2006 with some algorithmic advances according to Goodfellow et al. (2016); however, 2012 marks a more visible breakthrough of DL models. Krizhevsky et al. (2012) won the annual ImageNet contest (ILSVRC) using CNNs for image classification with a dramatically reduced error compared to earlier models. Since then, various competitions have been won, and different DL models

have achieved many unseen successes. Two of them attracted particular public attention because they were highly unexpected and came significantly earlier than anyone predicted. Both were achieved by the DL research team of Google (DeepMind). The first was in 2017 when DeepMind's *AlphaGo* beat the best Go-Players at the time (Silver et al., 2017), which was considered particularly difficult due to the large possibility of moves in the game and its highly complex strategies. The second breakthrough was DeepMind's *AlphaFold*, which already had won the protein folding contest CASP (Critical Assessment of Techniques for Protein Structure Prediction) in 2018. However, in 2020 *AlphaFold* tremendously outperformed earlier approaches and all competitors in such a way that some even considered the protein folding problem solved in some sense (Callaway, 2020; Jumper et al., 2021). Some important, and nowadays popular model architectures (e.g., LSTM) and algorithms originate from the '90s. The successes of DL in more recent times were then finally possible for the increasing amount of available data and the availability of computing resources to apply large models, such as training models on GPUs (graphics processing units), high-speed network connectivity, and infrastructure for distributed computing (Goodfellow et al., 2016).

2.2 Important Concepts and Prerequisites

The following section briefly introduces a few selected ML/ANN principles and concepts to provide an adequate context for the following chapters and better distinguish their contributions from another. However, the reader should already be familiar with ML basics such as learning algorithms, gradient-based optimization, over-/underfitting of models, hyperparameters, or regularization techniques. This list is not exhaustive, and it lies beyond the scope of this thesis to provide all fundamentals and prerequisites of ML and ANN modeling. Covering all these aspects would be worth several books on its own, and the interested reader is referred to renowned works such as Goodfellow et al. (2016), Bishop (2006), or Murphy (2012).

Time Series

Time series comprise data points that have a specific temporal order. Usually, time series exhibit autocorrelation, which means that successive points are related to each other (Hyndman and Athanasopoulos, 2021). On the one hand, this is an advantage because additional information is available for modeling besides the raw values of the data points. On the other hand, this also makes modeling much more complex since the temporal relationship within the data points and between different variables (e.g., input and output) needs to be considered.

Time series usually exhibit various properties. These include, for example, trend, seasonality, (irregular) cyclical behavior, or utterly irregular behavior (Hyndman and Athanasopoulos, 2021). Usually, ML methods can only handle stationary time series, i.e., constant mean and variance over time. Therefore, they must not show any trend, seasonality, or other cyclical behavior. This also applies to ANN modeling; however, if adequately designed and trained, ANNs can handle seasonality, especially if they contain an implicit or explicit memory to remember dependencies over time, such as RNNs do.

Basic Functioning of Neural Networks

The basic computation units of ANNs are called nodes, or neurons, in reference to their source of inspiration, the human nervous system. Neurons receive inputs from (i) one or several other neurons and (ii) a bias term. They compute their sum and apply a (mostly) nonlinear activation function. Typical nonlinear functions are displayed in Figure I.2, whereby the rectified linear unit (ReLU) is among the most popular functions currently. However, a wide range of modifications (such as leaky ReLU) and alternative activation functions exist.

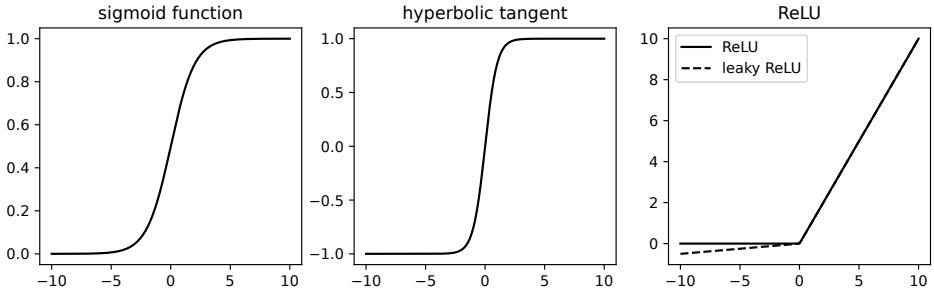


Figure I.2: Examples of nonlinear activation functions of neurons within ANNs.

Typically, neurons are ordered in layers (dense layers) and connected to the neurons in previous and subsequent layers but usually not to same-layer neurons. Each connection has a weight, which is modified during the learning process. A typical learning algorithm, s.a., *Backpropagation* (Rumelhart et al., 1986), applies the following simplified scheme in a supervised setting (see also subsequent section):

1. *Forward Pass*: propagate an input pattern through the network and compute an output.
2. Calculate error between output and ground truth.
3. *Backward Pass*: adjust connection weights and node biases to minimize the error.
4. Repeat

During the *backward pass* the algorithm computes derivatives to determine how much a model parameter needs to change in order to minimize the error. One can imagine this as a multidimensional error surface, where the algorithm computes the slope and seeks to step downwards to reach a (hopefully global) error minimum. We can decide how large the steps on that error surface are by adjusting a learning algorithm's (initial) learning rate. Choosing an appropriate learning rate is essential not to skip error minima (steps are too large), get stuck in local error minima (steps are too small) or slow down training unnecessarily on error surface plateaus (steps are too small). Therefore, the learning rate is usually adaptive, and the user chooses an appropriate initial value.

Supervised, Unsupervised, and Reinforcement Learning

ML research generally distinguished two fundamentally different learning techniques for a long time. Supervised learning (SL) uses labeled or target data to learn from. This means that ground truth is available to calculate and back-propagate an error during the model's training, which allows updating the model weights according to a loss function. In simple words, a teacher tells the model what's the right thing to do in a certain situation. Typical tasks based on SL are classification and regression. In contrast, unsupervised learning (UL) occurs in the absence of ground truth or human feedback. Here, the model learns to extract patterns from the data, solely using some kind of unsupervised criteria (e.g., compactness). However, the model does not know what to do with the found patterns and cannot evaluate their correctness. Usually, UL models aim to find a low(er)-dimensional representation of the input data, with possible applications ranging from density estimation, distribution sampling, or denoising to clustering (Goodfellow et al., 2016).

In 1998, Sutton and Barto (1998) introduced a third technique: reinforcement learning (RL). Models learn from being rewarded for correct decisions during trial and error actions. RL is argued to be the most similar to the way humans learn and has enabled most of the greatest successes in recent years, such as *AlphaGo*. RL is currently mostly limited to domains with vast amounts of training data (e.g., robotics) and is not part of this thesis.

Feedforward and Recurrent Models

Feedforward neural networks represent one of the earliest and most fundamental forms of ANNs. Many other architectures can be considered to be specific modifications or extensions of FFNs (e.g., CNNs) (Goodfellow et al., 2016). The most common form of FFNs is the multilayer perceptron. FFNs generally aim to find a function that maps any input to the corresponding output. Their name originates from how information flows through the net-

work, namely layer-wise forward, always from input to output layer (Figure I.3). FFNs can contain any number of hidden layers between the input and output layer, which is also the criterion mentioned above to distinguish between shallow and deep neural networks. They are further capable of tackling various modeling tasks, such as classification and regression. FFNs usually cope with temporal modeling using lagged or moving-window inputs to capture temporal dependencies between input and output variables.

Recurrent neural networks distinguish from FFNs by allowing information flow not only forward but also lateral and backward, depending on the exact structure of the RNN. These flow directions are achieved by implementing recurrent (or feedback) connections (Figure I.3), such that an RNN becomes internally aware of representations of previous time steps. This awareness means they possess a memory, which makes them particularly well suited for modeling temporal dependencies in time series. However, RNNs generally suffer from the vanishing and exploding gradient problems, which means that backpropagation learning fails over a larger number of time steps, and RNNs cannot remember long-term dependencies in the data (Bengio et al., 1994). A distinct RNN, however, which technically can remember information from an infinite number of time steps, is the LSTM, proposed by Hochreiter and Schmidhuber (1997). It applies gating mechanisms and an internal state to prevent information from vanishing. Another well-known model using this gating technique is the gated recurrent unit (GRU) introduced by Cho et al. (2014).

Both FFNs and RNNs are usually applied in a SL setting (as in this thesis) but can also be part of RL approaches. Figure I.3 illustrates the differences between FFNs and RNNs through visualization of the respective directions of information flow in both types. For the sake of simplicity, discrete connections between neurons are not drawn.

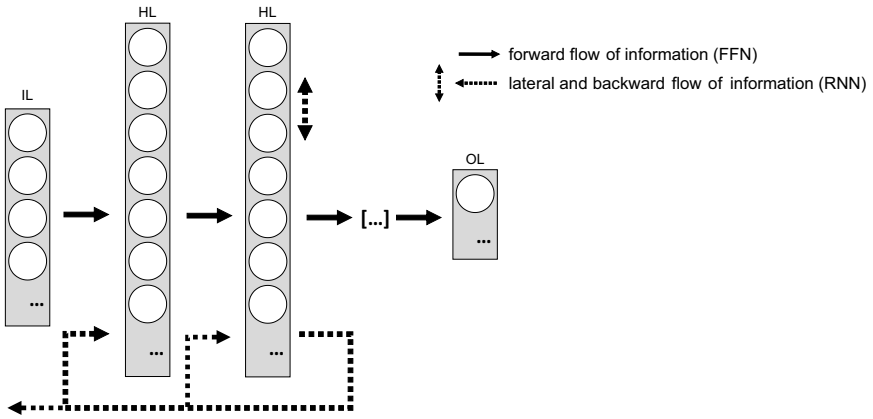


Figure I.3: Layered structure of a MLP with one input (IL), several hidden (HL) and one output layer (OL). Circles indicate single neurons, solid arrows illustrate forward flow of information in FFNs, dotted arrows illustrate possible additional flow directions in RNNs.

3 Outline

This thesis investigates different applications of ANNs, mainly in the groundwater domain, with a special focus on forecasting groundwater level time series on different time scales. In particular, the subsequent chapters will address the following research questions (RQs):

RQ1 How can we use unsupervised ANNs to group heterogeneous datasets of GW hydrographs based on their dynamics, and what can we learn from the resulting patterns?

RQ2 What are adequate model architectures to model and predict GWL time series, and what are their properties?

RQ3 Is it possible to perform reasonable short-term predictions of GWLs with ANNs without any future input data?

RQ4 What amount of data are necessary to build an ANN model for GWL prediction with reasonable performance?

RQ5 Can ANNs also be used to reasonably predict the long-term development of GWLs?

RQ6 How does the climate crisis influence the GWL development in Germany until the end of the century?

RQ7 How can state-of-the-art XAI techniques be used to increase trust in model decisions and to gain system understanding from ANNs models?

RQ8 How does a given routine for GWL modeling perform for predicting spring discharge in complex karst systems?

RQ9 Can ANNs learn the relevant fraction of spatially distributed input data automatically?

A total of four studies address these nine RQs. An elaborated approach for time series clustering with a specific adaption to heterogeneous GWL time series data sets investigates RQ1. Moreover, workflows for GWL prediction models are established, and the performance of several model architectures is comprehensively evaluated to examine RQs 2-4. To ensure high transferability, these workflows use solely easy-to-measure and widely available meteorological input variables, such as precipitation and temperature. Furthermore, the application of ANNs to long-term predictions (RQ5-7) and the successful transfer of GWL forecasting approaches to karst spring discharge modeling (RQ7-9) is demonstrated. See also Table I.1 for a summary at the end of this section. The following four paragraphs shortly present these applications, each of which corresponds to one study reproduced in the chapters II, III, IV

and V. Each paragraph summarizes the corresponding study, highlights the motivation, summarizes the main findings, and explains how all studies are connected. Publication details can be found at the beginning of each of the respective chapters. Finally, in chapter VI a synthesis of this thesis and an outlook are given.

The motivation of the first study in chapter II (Wunsch et al., 2022b) was to learn about factors that influence groundwater dynamics and to understand if, or how, these result in spatial patterns of (dis-)similar hydrographs (RQ1). For this purpose, an unsupervised clustering approach based on self-organizing maps (SOMs) (Figure I.4) was developed. SOMs are a powerful ANN approach with both characteristics of clustering (local averaging) and data compression methods (topology preservation) (Kohonen, 2014). However, most GWL time series are patchy and vary in length and covered period, making them inadequate for SOM and other clustering methods. Hence, direct processing of groundwater hydrographs was not possible or would have reduced the usable amount of data considerably. Surrogate clustering, which uses descriptors, also called features that capture certain aspects of groundwater dynamics, can overcome these problems. Moreover, such features can be calculated regardless of the quality of the primary data, at least to some degree. Some features were developed as part of this study, while others originate from the literature; in any case, they are all suited explicitly for describing the dynamics of groundwater hydrographs. This study shows that it is hard to conclusively assess and separate the influence of single factors on groundwater dynamics because they interact and superimpose both in time and space. Nevertheless, large groups of hydrographs with highly correlating dynamics, which are often also spatially grouped, were found. However, spatial proximity is no necessity for similar groundwater dynamics, and some patterns also emerge over larger distances. The application of this approach was demonstrated for the Upper Rhine Graben (URG) region in southwestern Germany/northeastern France.

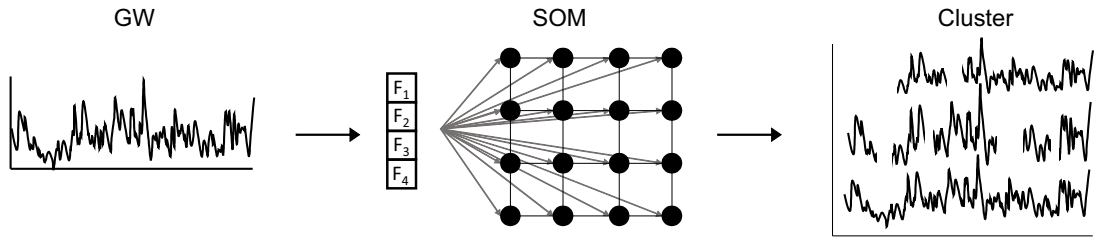


Figure I.4: Graphical abstract chapter II: Groundwater hydrographs are transformed into features and subsequently fed into a SOM model, which can cluster similar time series within a dataset.

Such groups of similar hydrograph dynamics are an excellent foundation for GWL forecasting because the results allow (i) to improve the individual data basis of a groundwater well by closing larger data gaps with information from highly correlated cluster neighbors, (ii) to

reduce the (computational) workload of forecasting a large dataset by selecting representative cluster members, and (iii) to conclude information on the representativeness of a single forecast for a larger region. Based on the results of this study, the data used in chapter III (Wunsch et al., 2021) is selected and preprocessed. Also, the data basis of chapter IV (Wunsch et al., 2022a) strongly profits by results of the approach developed here, even though applied in other regions than the URG.

At the time of performing the study presented in chapter III (Wunsch et al., 2021), a considerable increase of published modeling studies based on ANN and DL methods in the hydrological sciences occurred. Particularly LSTMs proved themselves extremely useful for several applications such as rainfall-runoff modeling (e.g., Kratzert et al., 2018) and increasingly became the method of choice for modeling hydro(geo)logical time series. In Wunsch et al. (2018), an earlier study, which is not part of this thesis, I already showed that nonlinear autoregressive networks with exogenous inputs (NARX) are very well suited to perform GWL predictions. However, focusing on popular DL methods such as LSTMs, pushed more classical ANNs (e.g., NARX) out of the focus of the studies conducted by the scientific community. The motivation of chapter III was therefore to compare these different model types specifically for the task of GWL forecasting, to find the best performing approach, and to learn about their properties (RQ2, RQ4) (Figure I.5). Besides NARX and LSTMs, also experiments with CNNs were conducted. At the time, CNNs were already successfully applied to signal modeling tasks in other than water-related domains such as NLP and had shown promising results in preliminary experiments for GWL prediction.

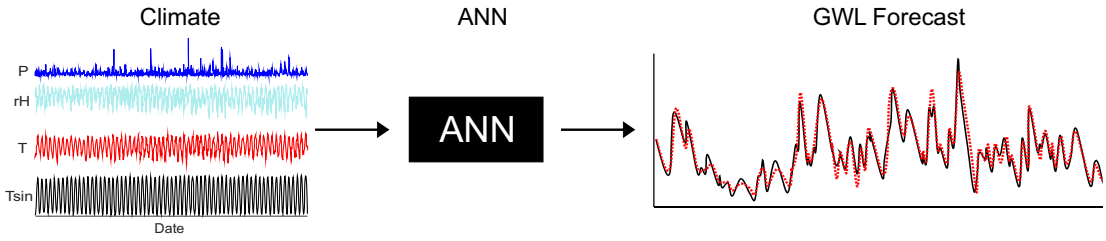


Figure I.5: Graphical abstract chapter III: Climate data serves as input to different ANNs to forecast groundwater levels.

Based on their properties and the results of existing studies, LSTMs, somewhat surprisingly, performed weaker than CNNs and NARX for a standard one-step-ahead modeling setting (sequence-to-value). NARX usually showed the highest performance values, closely followed by CNNs, while the latter exceeded both LSTMs and NARX substantially in calculation speed. Besides their high overall accuracy and speed, CNNs proved to be the most helpful tool for subsequent modeling studies because of the lower dependency on the random network initialization procedure and much greater implementation flexibility compared to NARX.

CNNs rely on a Python (van Rossum, 1995) implementation (instead of MATLAB as for NARX) using state-of-the-art frameworks such as Tensorflow and Keras (Abadi et al., 2015; Chollet, 2015).

Chapter IV (Wunsch et al., 2022a) was partly motivated by recent developments, such as the dry summers of 2018-2020 in Germany and their consequences, and is thus intended to contribute to answering questions about the direct influence of climate change on groundwater resources in Germany (RQ6) (Figure I.6). Another motivation was to investigate if ANNs could reasonably perform long-term predictions (RQ5), as it now was clear that ANNs perform excellently for short-term predictions. Such long-term predictions proved to be a particular challenge because it was necessary to ensure that the models learned the correct relationships and could reproduce them with high accuracy. Therefore, a way had to be found to establish as much confidence as possible in the simulation results, even for such long future periods, where no validation is possible (RQ7).

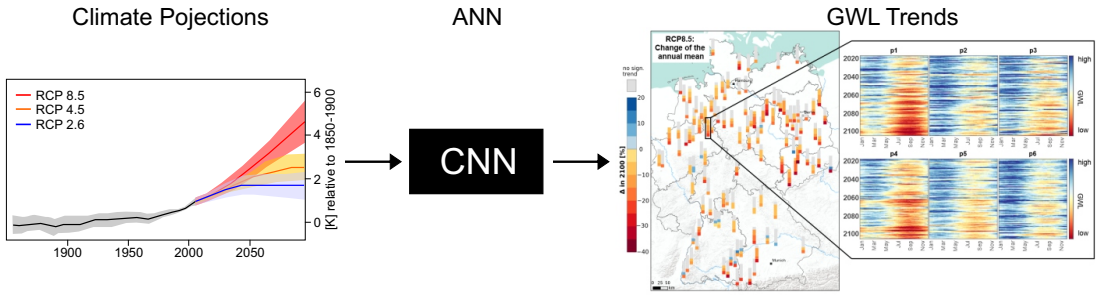


Figure I.6: Graphical abstract chapter IV: Climate data from different RCP scenarios are fed into CNN models to estimate the future groundwater level trends in Germany. Left part of the figure is based on (Tebaldi et al., 2021).

Climate projection data from three different representative concentration pathways (RCP2.6, 4.5, and 8.5) were the basis to simulate the future GWLs at 118 selected sites all over Germany. Each RCP describes one possible pathway of the future climate. The number-label of each scenario indicates the strength of the expected climate change and corresponds to the respective radiative forcing values in the year 2100. The simulations used purely climatic inputs; thus, they considered only the direct climate influence on groundwater. Future developments of other factors that strongly affect groundwater levels, such as anthropogenic extractions and changing land use or vegetation, had to be neglected due to missing data of their complex future development. A clear tendency of overall declining groundwater levels was found in the results for all scenarios, with partly opposite trends for annual upper extreme values (especially under RCP8.5), which illustrated the possibility of a generally increasing variability in the future for some regions and under certain conditions. Under the stringent mitigation scenario RCP2.6, the effects on groundwater are considerably less pronounced and less severe than for both other scenarios. Regardless of all political efforts in recent times,

the near future best matches the RCP8.5 conditions (Schwalm et al., 2020) and current estimations of future climate change impact still exceed the RCP4.5 scenario (UNFCCC, 2021), which highlights the importance of the results for RCPs 4.5 and 8.5.

The clustering approach from chapter II proved extremely useful in preliminary work to this third paper, to gain insight into dynamic patterns in regions of Germany other than the URG, and subsequently building the foundation to improve the data basis of thousands of time series that served as study site candidates for this study. Based on the findings of chapter III, CNNs were the method of choice to model the groundwater levels in this long-term study. The high framework flexibility allowed the implementation of an XAI approach that was key to selecting models and sites and increasing confidence in the simulation results. According to this XAI method, all models in this study learned the relation between input and output variables following the conceptual understanding of the major processes: groundwater recharge and evapotranspiration. Hence, these models not only performed well in the validation period but also "did the right things for the right reasons" without prior instructions or inductive biases.

Chapter V (Wunsch et al., 2022c) bridges the gap from GWLs to the closely related domain of karst spring discharge modeling (RQ8). Primarily due to the usually high complexity of karst systems, ANN approaches offer a convenient alternative to classical modeling approaches because only little karst domain knowledge is necessary to deploy them. The study results showed that CNNs are equally well suited to model karst spring discharge as to model groundwater time series (chapter III) and also rival existing modeling results of other authors in the studied areas. Motivated by the work of Anderson and Radić (2022), this study demonstrated that it is possible to let CNNs learn the relevant data from spatially distributed input data automatically (RQ9), which has the potential to solve data availability problems in many karst spring catchments by using openly available gridded meteorological data (such as E-OBS (Cornes et al., 2018)) (Figure I.7).

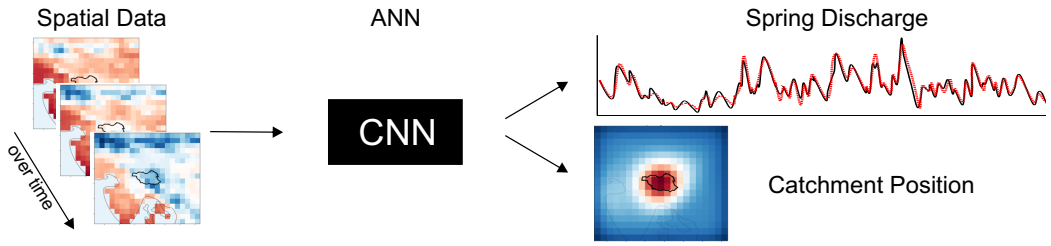


Figure I.7: Graphical abstract chapter V: Spatially distributed input data, fed into a CNN model allows both karst spring discharge simulations, and an estimation of the catchment position.

For this purpose, a combination of 2D- and 1D-CNNs processed data from three well-studied karst springs in Austria, Slovenia, and France, representing karst systems of different system properties, environmental conditions, and data availability. Furthermore, a spatial input sensitivity analysis of the trained models even opened possibilities of using this approach to localize karst spring catchments in the future (RQ7), given adequate conditions (such as the appropriate spatial resolution of the meteorological data) and further development of the existing approach.

Table I.1 finally provides an overview on the applied model types in the following chapters, how they relate to the concepts discussed in section 2.2, and which RQs are connected with each chapter.

Table I.1: Overview of applied learning concepts, models and the associated research questions.

Chapter	**Technique**		**Models**		**RQs**	**Goals**
	UL	SL	FFN	RNN		
II	x		SOM		1	Hydrograph clustering
III		x	CNN	NARX, LSTM	2, 3, 4	GWL forecasting, model comparison
IV		x	CNN		5, 6, 7	GWL long-term forecasting
V		x	1D-CNN 2D-1D-CNN		7, 8, 9	Karst spring discharge modeling

Chapter II

Hydrograph Clustering with Self-Organizing Maps

This chapter is based on a study published in Water Resources Management and combines the original article and material from its electronic supplement and is, therefore, a considerably extended version of:

Wunsch, A., Liesch, T., Broda, S., 2022. Feature-based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing Map-Ensembles. Water Resources Management 36, 39-54. doi: 10.1007/s11269-021-03006-y

The following links provide access to the associated online resources of this study:

Paper

Electronic Supplementary Material

Code

1 Introduction

The analysis and evaluation of groundwater level dynamics can contribute valuable information to assess quantitative groundwater availability, which is important to manage groundwater resources and secure water supply in many regions worldwide. As every hydrograph contains information about system properties (e.g., geology), artificial (e.g., withdrawal), and natural (e.g., streamflow interaction) environmental factors, hydrograph clustering is often helpful to identify common dynamics and to differentiate between signals resulting from external controlling factors and noise. This improves understanding of system dynamics and forms the basis for further analysis, including forecasting or scenario building. Popular methods to cluster hydrological time series are for example Cluster-Analysis (CA) (e.g., Naranjo-Fernández et al., 2020) and principal componant analysis (PCA) (e.g., Haaf and Barthel, 2018), each alone or as a combination of both (e.g., Machiwal and Singh, 2015). Besides classical approaches, ANNs offer innovative concepts to deal with larger sets of multidimensional data, such as using self-organizing maps for unsupervised clustering. Several studies from different disciplines compare SOM to other well-established clustering methods like k-means and hierarchical clustering (HC). Some authors found that k-means performs equally (He et al., 2004) or even better than SOM (Balakrishnan et al., 1994; Kumar and Dhamija, 2010; Mingoti and Lima, 2006); however, there is no consent on this aspect in literature as other authors found SOM to be clearly superior to k-means (Chen et al., 2010b; Kiang et al., 2006; Melo Riveros et al., 2019) and also to HC (Mangiameli et al., 1996). Often, SOM are even combined with k-means or HC methods because interpreting a trained SOM structure is not trivial, and usually, second-level clustering is therefore applied. Besides classical clustering methods, also algorithms specialized in the interpretation of trained SOM, such as DS2L (Cabanes et al., 2012), exist. In the hydrological context SOM have been extensively used to analyze water quality and chemistry (e.g., Gholami et al., 2021). Applications to groundwater hydrographs are: forecasting by using hybrid SOM-ANN models (Chang et al., 2016; Chang et al., 2014; Chen et al., 2010a; Lin and Chen, 2005; Moradkhani et al., 2004), hydrological event type clustering and classification (Abrahart and See, 2000; Toth, 2009), or catchment classification (Toth, 2013). The clustering of groundwater hydrographs, especially using SOM, has been carried out rather rarely so far. Han et al. (2016) used SOM to identify homogeneous clusters of groundwater level piezometers as a preprocessing step to forecasting with a step-wise cluster multi-site inference model. However, they tested the approach on a rather small number of wells (30), and more importantly, they used the time series directly as inputs. Such approaches that use time series directly for clustering suffer from dependency on high-quality data (equal length, equal period, no gaps). Application of feature-based approaches can overcome this problem by being able to use patchy input data (Wang et al., 2006). Features, in this case, are descriptive (statistical) measures

of the time series, extracted, e.g., from the time or frequency domain (Caiado et al., 2015). To apply a feature-based approach to groundwater level data, features taking the peculiarities of groundwater hydrographs into account are desirable. Heudorfer et al. (2019) present a comprehensive compilation of 45 possibly suited indices to describe groundwater dynamics. Their approach is very much related to the concept of hydrological signatures (e.g., McMillan et al., 2017), where features are designed to describe certain dynamic aspects in surface hydrology. Feature-based clustering of hydrological time series using self-organizing Maps has already been performed by Nourani et al. (2015), who used features based on wavelet decomposition to cluster a small number of wells on Ardabil plain, Iran. However, to the authors' best knowledge, no approach is known yet that combines SOM-clustering with specifically designed features that describe certain groundwater hydrographs' dynamics aspects.

In this study, we develop a robust, flexible, and semi-automated framework for groundwater hydrograph clustering. We deploy feature-based time series clustering, which allows us to use data from time series of different periods, different lengths, and missing and noisy data. Further, we present and explore several new features that showed promising results and are particularly suited to describe dynamic aspects of groundwater hydrographs. We modify a powerful clustering algorithm combination (SOM+DS2L) that allows influence on the level of detail of the clustering result and implement ensemble modeling techniques to remove arbitrariness from the feature selection process as to ensure higher robustness of the clustering result. We apply the developed approach to the Upper Rhine Graben area in central Europe, based on a dataset of overall 1853 groundwater hydrographs. The motivation and later application is the reduction of the forecasting workload of regional forecasting of groundwater levels by selecting representative hydrographs from the clustering result. Additionally, we aim for increased system understanding in terms of dynamic patterns and their main controlling factors.

2 Data and Study Area

2.1 Upper Rhine Graben Area

The study area is the Upper Rhine Graben, mainly located in southwestern Germany and northeastern France (Figure II.1a). It is the largest groundwater resource in central Europe (LUBW, 2006), covering 80% of the drinking water demand of the region (Région Alsace - Strasbourg, 1999) and is also intensively used for water extraction both for irrigation and industrial purposes. The URG, a Cenozoic rift structure, 300 km long (N-S) and on

average about 40 km wide (E-W), is filled with sediments (mainly gravel and sand) with a total thickness up to about 3500 m (Geyer et al., 2011). Hydrogeologically, the uppermost Quaternary sediments are most important. They reach a thickness of more than 200 m in the southern part, which strongly decreases to about 30 m in the area around Karlsruhe. In the northern part of the URG, the Quaternary sediment thickness increases up to 500 m, and a multi-aquifer system exists due to several fine-clastic layers dividing the Quaternary sediments (Geyer et al., 2011; LUBW, 2006).

2.2 Groundwater Data

The used dataset consists of 1853 weekly groundwater hydrographs from Germany and France, including one synthetic hydrograph with strong outlier characteristics to explore and illustrate additional properties of the clustering approach. The considered period ranges from October 1986 to September 2016 (30 years). The majority of the hydrographs contain data for almost the entire period; the shortest length being included is six years. We removed strong outliers conservatively and interpolated small data gaps up to 1 month linearly. Additionally, we set the maximum portion of data gaps within a time series to 25% and homogenized the dataset concerning the sampling interval. This included downsampling of higher-resolution data by picking discrete values (no averaging) and filling small data gaps to make use of monthly data. After preprocessing, considerable heterogeneity still exists, which was intended, since as much data as possible should be included, to fully use the available hydrographs. Figure II.1a shows the study area in general (left) and the locations of the 1852 actual wells included in the dataset (right). The dataset includes only wells from the uppermost aquifer within the Quaternary sediments, which causes, e.g., the three major blank spaces on the map in Figure II.1a (right), due to locally changing geological conditions in these areas.

2.3 Groundwater Dynamics

Figure II.1b sketches a strongly simplified E-W cross-section of the URG and illustrates that the regional groundwater dynamics are the result of a complex interaction of multiple factors, which we, for the sake of a more systematic point of view, divided into *processes* (*Pr*), *driving forces* (*DF*) and *governing parameters* (*GP*). *Processes* are the thereby physical processes that directly influence the groundwater levels (e.g., recharge). They are driven mainly by external *driving forces* (e.g., precipitation) and, in most cases, dependent on one or several *governing parameters* (e.g., topography, land use). The following paragraphs describe all factors in general and provide details on the respective conditions in the study area.

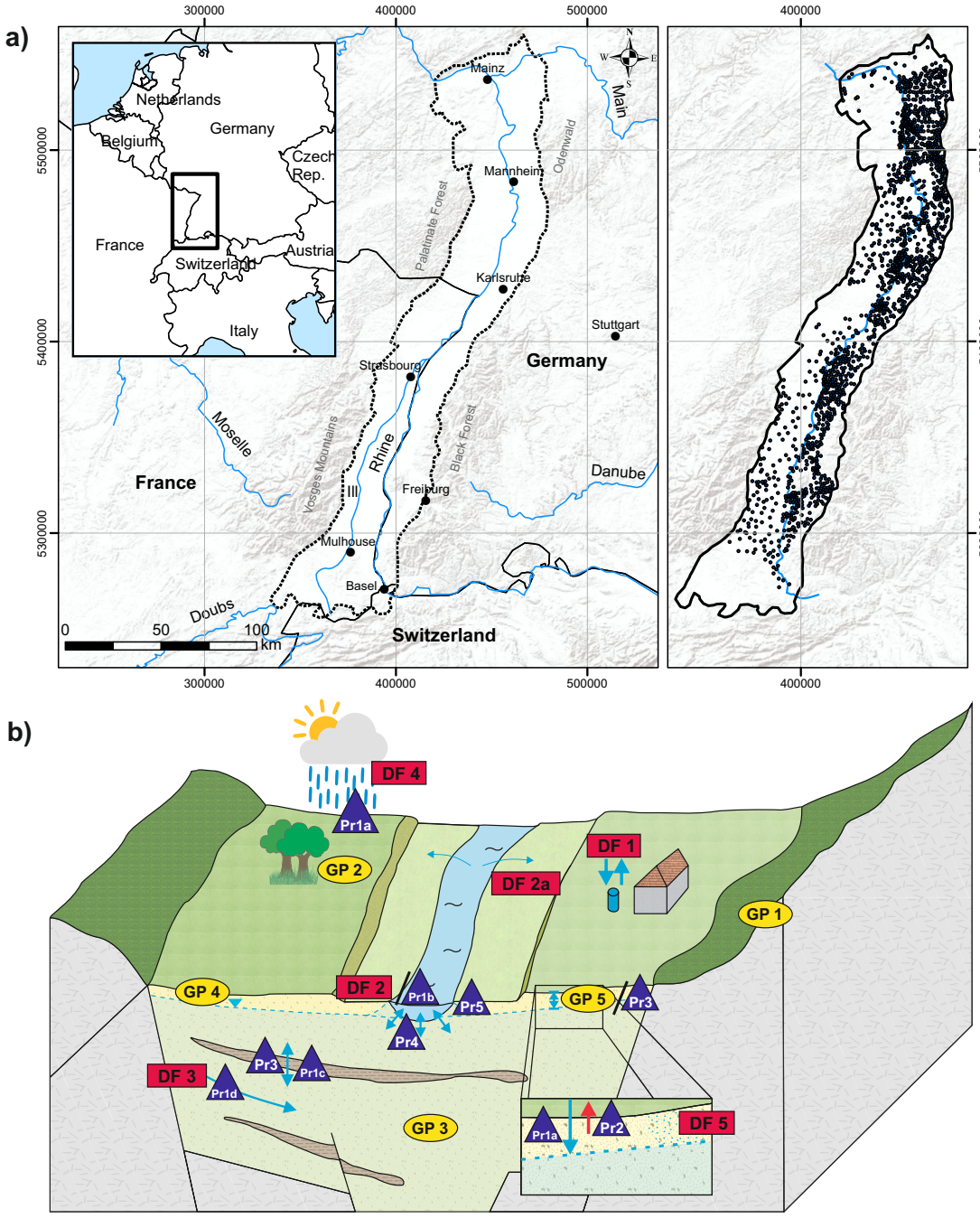


Figure II.1: **a)** The study area Upper Rhine Graben (left) and the locations of the used 1852 real groundwater monitoring wells (right). **b)** Strongly simplified E-W cross-section of the URG, summarizing some influences on groundwater dynamic patterns (*DF*: driving force, *GP*: governing parameter, *Pr*: process); *DF1* – artificial extraction/infiltration, *DF2* – surface water interactions (a: floods), *DF3* – regional flow systems, *DF4* – weather/climate, *DF5* – soil moisture; *GP1* – topography, *GP2* – vegetation/land use, *GP3* – geology (aquifer type/material properties), *GP4* – pressure state (free/confined), *GP5* – mean depth to groundwater; *Pr1* – recharge (a: direct/diffuse; b: direct/local; c: inter-aquifer-exchange; d: lateral), *Pr2* – evapotranspiration, *Pr3* – signal damping (low pass filter effect), *Pr4* – in-/exfiltration, *Pr5* – bank storage;

Groundwater Recharge and Climatic Conditions

One of the main processes with influence on groundwater dynamics in the region is groundwater recharge (*Pr1*), either directly (*Pr1a/b*) or as inter-aquifer exchange (*Pr1c*). Direct recharge is a highly complex process and occurs localized (*DF2/Pr1b*), or diffusely through the unsaturated zone (*Pr1a*). Recharge in general also depends on many other factors like precipitation (physical state, amount, intensity) (*DF4*), temperature (*DF4*), topography (*GP1*), vegetation (*GP2*), geology (*GP3*), soil moisture (*DF5*) etc. (e.g., Alley et al., 2002; Jasechko et al., 2014). Some of these, especially precipitation and temperature, in turn, are driven by global climatic patterns (*DF4*), which, especially in humid regions, have a significant influence on groundwater levels (Cuthbert, 2014) and generally influence factors like land use and vegetation (*GP2*) directly. These, in turn, have a strong impact on soil moisture (*DF5*) and evapotranspiration (*Pr2*). Further, mainly during long dry seasons, even in moderate climate shallow groundwater is exposed to the risk of strong direct groundwater evaporation (Balugani et al., 2017) as shown by Lam et al. (2011). The URG is one of the warmest areas in Germany, and the yearly precipitation within the Graben is in the order of 500 to 900 mm per year. The adjacent mountain regions can reach cumulative rainfalls of 2000 mm per year (Thierion et al., 2012). The mean annual groundwater recharge in our dataset ranges from 0 mm (mainly floodplains of the Rhine) to about 350 mm/a, with a mean value of about 150 mm/a. In general, the diffuse recharge in the northern part is comparably low, while the highest recharge values mostly occur in the middle URG between the cities of Offenburg and Rastatt (BGR, 2019). Dominant land use types within the URG are agricultural areas of different types (37%), on par with artificial surfaces (36%), the rest are mostly forests/semi-natural surfaces (22%) (CORINE Land Cover, 2018).

Hydrogeological Properties and Regional Context

Geology (*GP3*), thus, material properties (permeability/hydraulic conductivity, effective porosity) or, more generally speaking, the aquifer type (porous, fractured, karstic), also plays a major role in controlling groundwater dynamics. Porous unconsolidated gravel or sand aquifers like in the URG usually show high matrix porosities, often going along with high hydraulic conductivity and high storage capacity. Also, the regional geological setting is of great importance, since the development of local and regional groundwater flow systems (*DF3*), thus the lateral recharge (*Pr1d*) within an aquifer, depends on it (Toth, 2009). Confined and unconfined aquifers (*GP4*) are known to react differently to atmospheric pressure changes or groundwater withdrawal (Alley et al., 2002; Hölting and Coldewey, 2013). The mean depth to groundwater (*GP5*) is also an important factor concerning groundwater dynamics as the recharge signal is increasingly damped with depth (*Pr3*), filtering seasonal

variation patterns and leaving only multi-annual periodicities. Overlying layers with lower hydraulic conductivities can amplify this low-pass filter effect (e.g., Corona et al., 2018). The study area comprises mainly unconfined sand/gravel aquifers of generally high storage coefficients and high hydraulic conductivities in the order of 10E-4 to 10E-3 m/s (LUBW, 2006). Hydrographs used in this study are from the uppermost aquifer, with very shallow mean depths to groundwater (<5 m bgl for 70% of the wells), rising to a maximum of about 20-30 m towards the Graben edges. A rather shallow gradient towards the north of the Graben and at the same time from the Graben edges towards the graben center controls the regional groundwater flow-systems (Thierion et al., 2012). Towards the Graben edge, local inflow from adjusting fissured aquifers or alluvial fans from side valleys may dominate the flow regime and result in steeper gradients towards the Rhine River as the main receiving streamflow of the region.

Surface Water

Surface water interactions (*DF2*), already mentioned as a source of local recharge, are usually essential driving forces of groundwater dynamics. Important processes and driving forces in this context are, for example, streamflow in-, and exfiltration (*Pr4*), bank storage (*Pr5*), tides, waves, as well as floods (*DF2a*)(Alley et al., 2002; Cloutier et al., 2014). In the study area, the main surface water body is the Rhine River, with a strong influence on groundwater dynamics, up to several hundreds of meters in distance. To a lesser degree, there are also smaller streams from the adjacent mountain ranges that strongly affect groundwater dynamics on a local scale (Longuevergne et al., 2007). Besides natural interaction, especially in floodplains and along the ancient river course, anthropogenic interventions like correction of the streambed course or weir locks and dams influence the dynamics in many parts along the streams.

Artificial Factors

Anthropogenic actions, in general, cannot only influence streamflows but also strongly alter groundwater dynamics directly (Stoll et al., 2011). Typical influences in general, also widely present in the study area, are land-use changes over vast areas, landscape-engineering actions (e.g., river course modifications and dredging lakes), recharge inhibition by surface sealing in urban areas, abstraction for drinking water supply or industrial purposes, artificial infiltration, and irrigation in agricultural areas, which increased in the study area particularly in recent years. Especially direct groundwater interactions like abstractions and infiltrations (*DF1*) are most important because, on a local scale, pumping patterns can partly or even completely

superimpose the natural groundwater dynamics. Especially in the northern part of the URG, intensive groundwater management is applied by managing extraction rates and artificial aquifer recharge. Besides the increasing water demand in these areas, this is especially necessary to protect ecosystems and infrastructure from land-subsidence and groundwater-floodings (Bouwer, 2002; Regierungspräsidium Darmstadt, 1999).

3 Methodology

3.1 Feature-based Time Series Characterization

Depending on the unique hydrogeological conditions, a proper feature set is a key to adequately describing and thus successfully clustering the data. Here, features are descriptive (statistical) indices that quantify the dynamics of groundwater hydrographs, similar to the concept of signatures in hydrology (see, e.g., McMillan et al., 2017). However, groundwater hydrographs generally differ considerably from surface water hydrographs, which makes many hydrological signatures inadequate for describing dynamic aspects of groundwater, and there is a need for comprehensive testing of transferability to the groundwater domain like done by Heudorfer et al. (2019). The most important supportive tool for pre-selecting adequate features is a visual skill test to check every single feature's adequacy and explanatory power. Applying PCA or related methods can help to reduce the feature number by ruling out redundant features based on the explained variance. However, including correlated features can help to improve the result by up-weighting important aspects of the general dynamics. We explore this aspect with a correlation analysis of all selected features in the results section. In total, we tested a broad variety of feature candidates (>50), including standard statistics measures, features derived from literature (e.g., from Heudorfer et al., 2019; Wang et al., 2006), as well as self-designed features to account for peculiarities of both the study area and groundwater hydrographs in general. In the following, we introduce those that have successfully passed the visual skill test for our data set. Skill test results that show the explanatory power of each feature are provided in the supplementary material (Figures S1 to S13). Table II.1 summarizes the feature calculation, the corresponding data basis, and the primary purpose or a short description for all used features. For more details on the self-designed features, we refer to the supplementary material (Text S2).

We designed three experiments to examine better the properties and data requirements of the applied features. The first two try to answer how strongly the features react to missing values and white noise in the data, respectively. Thus, 0%-25% of each time series is randomly replaced by white noise or data gaps in 0.25% steps (50 times each), and both the

Table II.1: List of promising features (passed skill test) to describe groundwater dynamics of time series in the URG dataset. Features in italic were not used based on the decision of the ensemble approach (see section 3.3)

Feature Name (Abbrev.)	Data*	Purpose / Description	Ref**
Range Ratio (RR)	o	Detection of superimposed long-periodic signals, also sensitive to outliers, calculated as the ratio of the mean annual range to the overall range	sd
Skewness (Skew)	o	Boundedness, inhomogeneities, outliers, asymmetry of the probability distribution	ss
Annual Periodicity (P52)	o	Strength of the annual cycle, calculated by correlating (Pearson) the mean annual (52 weeks) periodicity with the complete time series	sd
SDdiff	o	Flashiness, frequency and rapidity of short-term changes, calculated as the standard deviation of all first derivatives	sd
Longest Recession (LRec)	o	(unnaturally) long descending heads, longest sequence without rising head values	sd
Jumps	z	Inhomogeneities/breaks, partly also variability, calculated as the absolute and standardized maximum change of the mean of two successive years	sd
Seasonal Behavior (SB)	z	Position of the maximum in the annual cycle, agreement with the expected average seasonality (Min in September, Max in March)	sd
Median[0,1] (Med01)	n	Boundedness, median after scaling to [0,1], standard statistics measure, derived from (Heudorfer et al., 2019)	ss/lit
High Pulse Duration (HPD)	n	Average duration of heads exceeding the 80$^{\text{th}}$ percentile of non-exceedance, for details see Richter et al. (1996), derived from (Heudorfer et al., 2019)	lit
Richards-Baker Index (RBI)	*o*	*Flashiness, frequency and rapidity of short term changes, for detailed explanation see Baker et al. (2004)*	*lit*
Yearly Variance (Yvar)	*z*	*Variability, periodicity, calculated as the median of the yearly calculated variances*	*sd*
Standard Error of the Mean (SEM)	*o*	*Standardized statistical dispersion, calculated as the standardized standard deviation of the time series*	*ss*
Low Pulse Duration (LPD)	*n*	*Average duration of heads dropping below the 20th percentile of non-exceedance, for details see Richter et al. (1996), derived from Heudorfer et al. (2019)*	*lit*

** o: original, z: z-scored, n: normalized*
*** lit: literature, sd: self-designed, ss: standard statistics*

absolute characteristic values and their changes compared to the initial undisturbed values are examined. To estimate how long a time series has to be at least to provide a representative feature value, the features for systematically varied time series lengths were calculated in experiment three. Starting from 2016, the time series length was extended in 1-year steps until 1986. For this experiment, we used only a subgroup of about 50% of the data set, which had complete data over the 30 years. To make the feature values and changes comparable, the features were standardized in all experiments, using the respective mean and standard deviation from the (undisturbed) 30-year feature values.

3.2 Self-Organizing Map Clustering Using DS2L Algorithm

SOM perform a nonlinear projection of multidimensional data onto a regular neuron lattice surface. They show characteristics of both clustering (local averaging) and data compression methods (topology preservation), which is a unique property and also an advantage of SOM compared to other cluster algorithms and projection methods (Kohonen, 2014). Every neuron has clearly identifiable neighbors, allowing simple two-dimensional visual representations of multi-dimensional data. We apply a modified version of the density-based simultaneous two-level (DS2L)-algorithm (Cabanes et al., 2012) to automatically derive clusters from the trained SOM. DS2L detects clusters by analyzing data density and neighborhood connection strength of the SOM. An adequate cluster number is automatically determined, and the algorithm does not tend to produce clusters of equal size, both advantages compared to some well-established cluster algorithms (e.g., k-means or some hierarchical methods). We modify DS2L-algorithm so that the user can decide purely qualitatively whether the clustering should be performed more coarsely or more finely. The cluster number is still determined automatically on the chosen level of detail. For this, we implement three adjustment parameters for thresholds of data density and neighborhood connection strength as well as to control the application of some algorithm steps. Besides the number of neurons (SOM-size), which also influences the clustering result, the following four parameters must be optimized during the clustering process.

- **SOM-size**: *normal* ($5\sqrt{n}$), *small* ($5\sqrt{n} \cdot 0.25$) or *big* ($5\sqrt{n} \cdot 4$) - options implemented in SOM-Toolbox (Vesanto, 2005), n: number of samples
- **NTH**: $NTH \geq -1 \in \mathbb{Z}$ - DS2L-Neighborhood-Threshold, connection strength required to qualify as cluster border, -1 means connection strength is not used.
- **DR**: Yes/No - DS2L-Density-Refinement, use density values for cluster determination
- **DM**: Yes/No - DS2L-Density-Merging, merge similar clusters based on density-dependent index

3.3 Workflow

Figure II.2 summarizes the workflow of the approach applied in this study. A common problem with many feature-based approaches is the arbitrariness of feature selection. As shown by line I in Figure II.2 we implement a SOM-ensemble to find the best combination of all pre-selected features, whereby the cluster quality is judged by five different internal validation indices (Caliński-Harabasz criterion (CH), McClain-Rao criterion (MR), PBM-Index, Ratkowsky-Lance criterion (RL), C-Index). Line II in Figure II.2 shows a second SOM-Ensemble based on delete-d-Jackknifing resampling. Its purpose is to simulate changes in the observational network by manipulating the input data set and to obtain cluster results as robust as possible. The final cluster result is based on voting consensus. We rearrange all original time series of a cluster for visualization and evaluation by their mean pairwise Pearson correlation with all other cluster members. A weighting by the p-value of the respective single correlations lowers correlation values with low significance (which might arise from only short overlapping periods). We define this value as the *weighted intra-cluster correlation* ($\overline{R_W}$). A detailed description and discussion of the workflow is added to the supplementary material (Text S3).

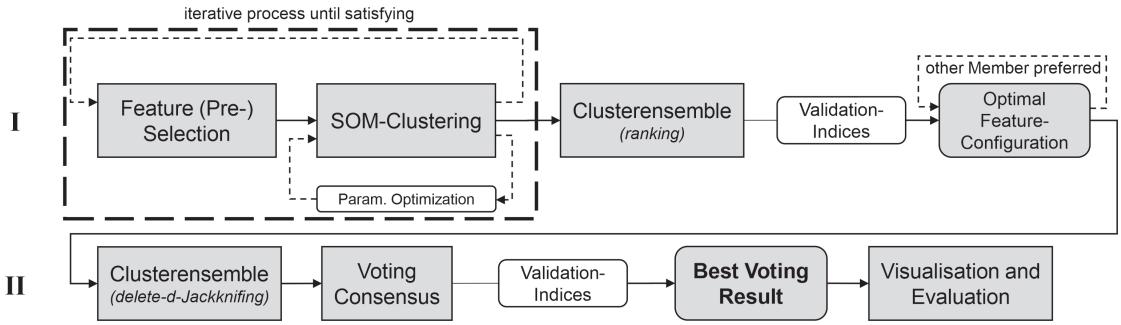


Figure II.2: Workflow of the presented methodology.

Besides the clustering itself, interpreting the results is very useful to improve system understanding in general. This is especially the case for clusters, which are not easily interpretable in terms of spatial location or dynamic aspects. Hence, we conduct detailed correlation analyses for factors mentioned in Figure II.1b, where reasonable additional data are available to perform meaningful statistics. For some, data are only available for a part of the study area; we, therefore, link them also with features and not only with clusters. In this way, we avoid a bias for clusters with wells in areas without data. Furthermore, the dynamics within clusters are usually the result of a superposition of several influencing factors, which can make correlations rather challenging. We focus on linear correlation analysis because of the easier metric interpretation, although we are aware that nonlinear relationships can also exist; further, we only mention significant correlations with $p<0.05$.

4 Results and Discussion

4.1 Feature Robustness

We conducted three experiments to examine the sensitivity of the 13 tested features towards (i) data gaps, (ii) white noise, and (iii) time series length. These experiments mainly aimed to test if thresholds exist, which can be formulated as recommendations for minimum data quality requirements. As reference values within each experiment, the undisturbed values of the features (no additional gaps, noise, or shortened time series) were used. We can show (Table II.2) that most features only react little ($<$0.1 with 25% missing values) to additional data gaps. In contrast, adding white noise leads to much higher differences much faster. Though one might think this could lead to unstable results for noisy datasets, this is probably not the case in reality. Little noise from unknown sources is hard to recognize and will not lead to strong differences in feature calculation. However, strong noise causes higher differences in the features, usually can be detected as outliers and removed hereafter. Therefore, data should always be carefully checked for implausible outliers in preprocessing.

Experiment three shows that time series length seems to have a constant influence on the feature values (Table II.2). We found a steady increase of differences the shorter the time series, up to strong increases for lengths of only a few years. No threshold value that might serve as a recommendation as minimum length can be found. Thus, we instead generally conclude that the longer the time series, the better. Features that are not robust and show bad performance, or cause unsatisfying cluster results, should usually be ruled out by the visual skill test or the feature selection ensemble. Please check the supplement for detailed information on feature robustness results (Tables S2 and S3). Answering how these disturbances alter the clustering result is exceptionally challenging because additional factors such as the ensemble and the consensus voting approaches also influence the final results. Extensive and thorough experiments would be necessary to investigate these interactions, which is why this question lies beyond the scope of this work but would be worth to be answered in future research.

4.2 Clustering Results

We applied our approach to 1853 time series from the URG (incl. one synthetic hydrograph). The feature pre-selection provided 13 features with good explanatory power regarding our specific dataset (sec. 3.1/Table II.1). The used cluster parameter combination was: *SOM-size*: big, *NTH*=0, *DR*=Yes, *DM*=No (sec. 3.2). The best feature configuration derived from the first ensemble (115.005 members) included 9 out of 13 features.

Table II.2: Median influences of data gaps, white noise and time series length on standardized feature values. The table shows the absolute values of the differences between the according disturbed values and the undisturbed values (no additional noise, data gaps, full length). Due to the standardization, the unit below is standard deviations.

	Added **White Noise**				Added **Data Gaps**			
Feature	1%	5%	10%	25%	1%	5%	10%	25%
RR	0.13	0.49	0.75	1.14	0.00	0.01	0.02	0.04
Skew	0.00	0.03	0.05	0.12	0.00	0.00	0.00	0.01
P52	0.02	0.08	0.15	0.34	0.00	0.01	0.03	0.07
SDdiff	0.41	1.37	2.10	3.42	0.01	0.03	0.06	0.15
LRec	0.00	0.20	0.40	0.69	0.00	0.06	0.11	0.23
Jumps	0.01	0.03	0.05	0.11	0.00	0.02	0.03	0.12
SB	0.00	0.02	0.03	0.08	0.00	0.00	0.01	0.01
Med01	0.00	0.11	0.30	0.49	0.00	0.00	0.00	0.00
HPD	0.10	0.31	0.42	0.52	0.01	0.03	0.04	0.10
RBI	*0.03*	*0.24*	*0.46*	*0.99*	*0.00*	*0.01*	*0.01*	*0.03*
Yvar	*0.00*	*0.00*	*0.00*	*0.01*	*0.01*	*0.04*	*0.09*	*0.22*
SEM	*0.08*	*0.27*	*0.38*	*0.48*	*0.01*	*0.03*	*0.04*	*0.10*
LPD	*0.14*	*0.68*	*1.31*	*2.82*	*0.01*	*0.04*	*0.08*	*0.18*
	Time Series Length [years]							
Feature	30	25	20	15	10	5	3	1
RR	0.00	0.05	0.24	0.40	0.87	1.47	2.10	NaN
Skew	0.00	0.08	0.17	0.29	0.27	0.39	0.53	0.51
P52	0.00	0.07	0.16	0.18	0.45	0.78	0.78	2.68
SDdiff	0.00	0.02	0.04	0.06	0.09	0.10	0.14	0.16
LRec	0.00	0.00	0.00	0.00	0.06	0.34	0.40	0.92
Jumps	0.00	0.14	0.23	0.44	0.91	1.39	1.76	NaN
SB	0.00	0.04	0.07	0.12	0.16	0.22	0.32	0.61
Med01	0.00	0.10	0.29	0.46	0.53	0.60	0.85	0.95
HPD	0.00	0.05	0.08	0.10	0.14	0.15	0.17	0.31
RBI	*0.00*	*0.06*	*0.15*	*0.21*	*0.51*	*0.76*	*1.38*	*NaN*
Yvar	*0.00*	*0.15*	*0.26*	*0.47*	*0.60*	*1.49*	*2.13*	*5.08*
SEM	*0.00*	*0.03*	*0.05*	*0.08*	*0.12*	*0.17*	*0.19*	*0.27*
LPD	*0.00*	*0.02*	*0.03*	*0.05*	*0.07*	*0.09*	*0.10*	*0.13*

As stated in section 3.1, we found that including correlated features improves the clustering results. A correlation analysis among the included features shows the highest absolute significant ($p<0.05$) correlations for the features Skew-Med01 (-0.81) and P52-RR (0.79), which is consistent with the meaning and calculation of these respective feature pairs (e.g., hydrographs with high annual periodicity often also show a regular range over the years, thus high RR values). A detailed correlation matrix of all features can be found in the supplementary material (Figure S27).

The final cluster result consists of 18 clusters (Figure II.3a) with sizes ranging from 239 hydrographs in cluster 1, to only one hydrograph in cluster 18, which is the synthetic hydrograph with outlier characteristics (cluster numbers sorted in descending order by size).

The five biggest clusters include almost 1000 of the 1853 hydrographs in total; 8 clusters show sizes larger than 100, only 5 clusters show sizes below 50. Due to the vast amount of information, we summarize detailed information and graphics on every single cluster in the supplement (Figures S28-S65); we only present selected results in the following.

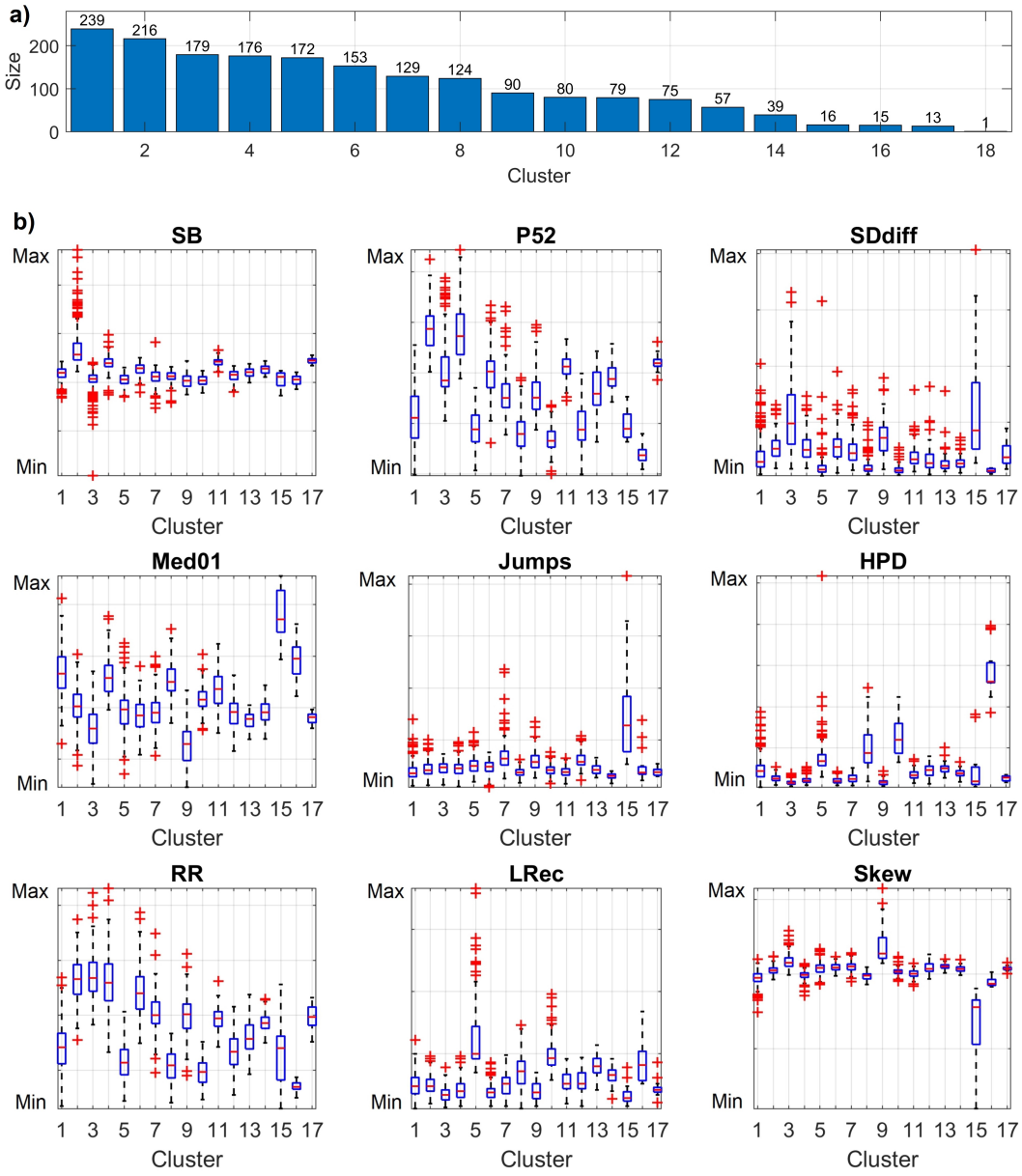


Figure II.3: **a)** Cluster sizes; **b)** Feature-value boxplots of all clusters. For a better graphical representation Cluster 18, was omitted, due to strong outlier characteristics. Boxplots including Cluster 18 can be found in the supplement.

The boxplots in Figure II.3b show the feature value distributions within each cluster. For some clusters, clear feature importance can be derived. Cluster 2, for example, is comprised of mainly regular hydrographs dominated by annual periodicity and with few other long- or short-term periodicities (high P52), as well as the annual maximum and minimum occurring very regularly during March and September respectively (high SB). The reasons are comparably high recharge values in the middle of the Graben, which are typical for wells neither strongly dominated by margin inflows nor by the Rhine River. However, less straightforward feature combinations also exist, which are harder to interpret. The same applies to the spatial distribution of the clusters. Without distinct grouping, e.g., as a result of a spatially limited, local influence on the dynamics, more effort is required to understand what processes, forces, or parameters might cause the common dynamics.

Cluster 3 (Figure II.4a) is an example of straightforward interpretation, where wells follow almost exclusively the Rhine River course, thus identifying interaction with surface water (*DF2*, *Pr1b*, *Pr4*, *Pr5*, Figure II.1b) as the dominant driving force is comparably easy. Some wells of this cluster showing greater distances to the Rhine River are in turn closer to mid-sized rivers like the Neckar or Ill, where common dynamics can be expected due to similar overall conditions. The resulting hydrograph grouping reveals that despite data gaps and different time series lengths, the homogeneity of the cluster is still high. The weighted intra-cluster correlation values ($\overline{R_W}$) are expressed by the coloring (the brighter, the lower), thus by the sorting of the stacked time series and by the bars on the right. In general, with decreasing ($\overline{R_W}$)-values towards the cluster borders, the heterogeneity increases, and the certainty of the cluster assignment of individual hydrographs decreases. Considering cluster 3, we can observe a distinct north-south gradient, which means that despite a changing dynamic along the river, grouping was still successful. Other wells close to the Rhine River were sorted into different clusters but showed indeed different dynamics (compare clusters 7 and 9 in the supplement). In terms of feature values, the Rhine influence for cluster 3 is best expressed by feature SDdiff, describing the higher flashiness close to the river (Figure II.3). Other features are also in accordance. For example, Med01 values are comparably low, indicating that the hydrographs are more likely to be bound to some kind of baseflow level in combination with short and high peaks triggered by the streamflow.

Overall results show that in the north of the URG predominantly hydrographs with small variability and weak annual periodicity occur, while especially the middle section of the URG exhibits highly seasonal and highly regular hydrograph patterns. The former is expressed mainly by clusters 1, 5, 8, 10, 16; the latter can be seen, e.g., in clusters 2 and 4 (Figures S29-S65). We selected cluster 8 (Figure II.4b) to illustrate the low-variance case in the northern URG. Driving forces connected to this cluster are most certainly strong anthropogenic influences (*DF1*, Figure II.1b) because the cluster focuses spatially on an area with strong

groundwater management efforts. Connections to generally lower groundwater recharge values (*Pr1a*, Figure II.1b) in the northern URG can also be drawn. Both factors can explain the smoothness as well as the comparably weak annual periodicity and low variability of the hydrographs in cluster 8.

The approach successfully separates a small group of 16 hydrographs with outliers and significant inhomogeneities, which probably occur due to two major Rhine River weir locks (Strasbourg, Breisach) (cluster 15, Figure II.4c). Furthermore, the synthetic hydrograph is put in a separate cluster (cluster 18, Figure II.4d). Both clusters are examples for clusters that are rather based on single events or characteristics than on similar, highly correlated time series. Therefore, even for good clusters in terms of such events, ($\overline{R_W}$)-values can be rather low.

In terms of system understanding, thus the correlation analysis of clusters and features with explaining factors, we found that the mean depth to groundwater (*GP5*, Figure II.1b) shows clear negative correlations (P52 (-0.45), RR (-0.44), SB (-0.29), SDdiff (-0.16)) with features describing the variability of hydrographs (e.g., seasonality, flashiness). Such variability is generally damped with increasing depth to groundwater. The complimentary case applies for HPD (0.33) and LRec (0.29), which reach higher values for smoother hydrographs with little short-term variations. A clear relation to the clusters could not be found, though, probably due to the just minor variation of this parameter (70% of the wells $<$5 m bgl on average), which makes a meaningful interpretation of the cluster development challenging. We observed only slight tendencies to greater or smaller depths to GW for some clusters. Another probable explanation could be that more dominating factors superimpose the effect of the depth to groundwater and are thus more decisive for cluster assignment.

We explored the connection of features and clusters to diffuse groundwater recharge (*Pr1a*, Figure II.1b) using the GWN1000 dataset (BGR, 2019). French wells (190) were excluded due to no data. In accordance with the findings and explanations given for depth to groundwater, we found significant positive correlations for damping sensitive features (RR (0.26), P52 (0.19), SB (0.07), SDdiff (0.05)). Further, it seems plausible that weak recharge signals correlate with important features for smoother hydrographs (LRec (-0.15), HPD (-0.14)). In agreement with the spatial recharge data, we found that clusters showing mainly smooth hydrographs with lower variability (1, 5, 8, 10, 16) are connected to lower recharge in the northern URG; clusters showing higher annual periodicity and variability and which occur mainly in the middle part of the URG (2, 4, 6) are connected to higher recharge. Nonetheless, due to missing data for France, these relations must be considered somewhat carefully.

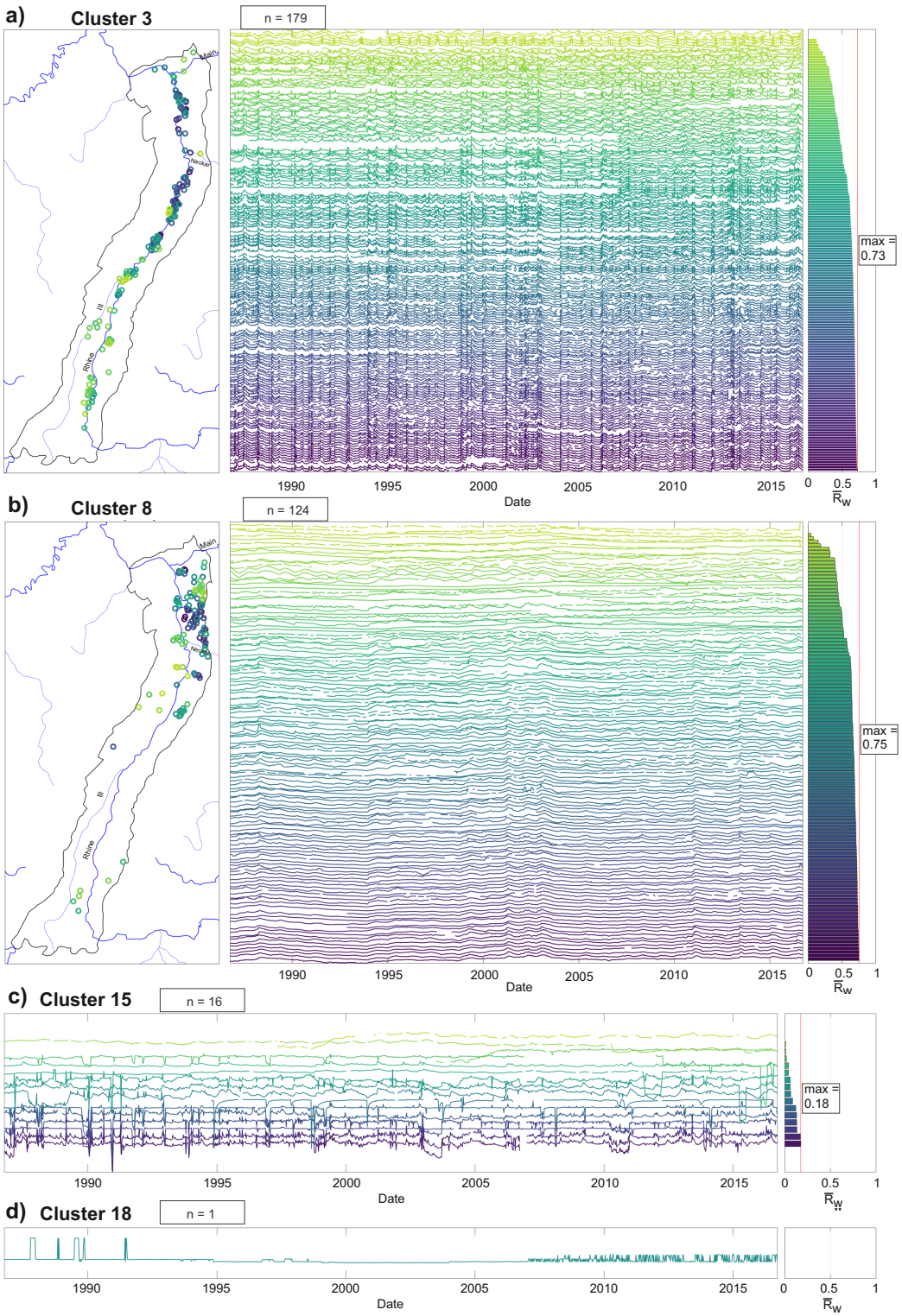


Figure II.4: Maps and stacked, z-transformed hydrographs of selected clusters. Coloring and stacked order reflect the weighted intra-cluster correlation ($\overline{R_W}$), also shown as bar-plot on the far right; **a)** Cluster 3 is mainly influenced by the Rhine River; **b)** Cluster 8 shows spatial grouping in the northern part and contains hydrographs with low annual periodicity and low variability; **c)** Cluster 15 groups hydrographs with outliers and inhomogeneities; **d)** Cluster 18 contains only the synthetic hydrograph, which is a heavy outlier compared to the whole dataset.

For most of the area east of the Rhine (Baden-Württemberg) we explored connections to the hydraulic conductivity within the uppermost aquifer (K-values, *GP3*, Figure II.1b) (LGRB, 2007). Due to the spatially limited data, no meaningful correlation can be made with clusters; however, a reasonable number of wells (828) can still be assigned to a specific K-value. Categorical correlation analysis (Spearman) with features yields positive correlations for Skew (0.24) and SDdiff (0.18), probably because high conductivities can be found mainly close to the Rhine River. Similarly, Jumps (0.20) are probably often caused by anthropogenic influences (GW abstractions, ship locks), which in turn occur preferentially in regions of high conductivities. Other correlations implicate that smoother hydrographs (HPD (-0.34)), long descending hydrograph parts (LRec (-0.23)), boundedness preferentially to an upper bound (Med01 (-0.21)), as well as the yearly maximum during spring (SB (-0.18)), seem to be related to lower hydraulic conductivities for this subset of wells. This might sound counter-intuitive since flashy behavior is often linked to lower hydraulic conductivities, however, the main reason for flashy behavior in this area is probably the influence of the Rhine River, where high conductivities occur.

The influence of streamflows (*DF2*, *Pr1b/4/5*, Figure II.1b) was first explored as the general relationship between distance to the Rhine River and feature values. The results confirm the relation to cluster 3. Further, we found clear relationships for clusters 7 and 9. Clusters 6 and 15 showed a weaker connection, but all of the mentioned clusters show a clear spatial relation to the Rhine River. Nevertheless, they exhibit different dynamics, which maintains the reasonability of the results. Clusters 3, 6, and 7 are closely related, but the hydrographs' flashiness decreases from one to the other. Cluster 9 shows fewer periodicity than cluster 3, but both are visually similar and match major dynamic peaks. It remains an open question what the causes of different dynamics close to the streamflow are. Also, smaller streamflows seem to have a significant influence on groundwater, at least in the southern part of our test area (Longuevergne et al., 2007). Hence, secondly, we performed a detailed streamflow distance analysis based on the Strahler classes of all streams (Text S5) within the area, derived from the Copernicus *EU-HYDRO* Dataset (EEA, 2017). We obtained similar findings, showing a stronger influence for cluster 15 and a slight influence of streamflows on cluster 12.

For most conducted analyses, the correlation values are significant, but rather low. This illustrates that there are distinct relations but at the same time also a lot of interactions between the influences. Correlation is nevertheless a good indicator and shows that the features express important properties of the hydrographs and thus are well selected. On the other hand, low correlations also indicate that dynamic-based clustering is even more important because simply grouping wells according to external factors is insufficient. Supplement Table S4 and Figure S66 show a comprehensive overview of all explored correlations (r-values and significance).

5 Summary and Conclusions

In this work, we present the results of a newly developed semi-automated groundwater hydrograph clustering framework. We group hydrographs based exclusively on their dynamics by describing them with features designed explicitly for important dynamic aspects of groundwater hydrographs. Heterogeneous input data can be used, which we confirmed by high robustness for most of our features, especially towards data gaps. Combining the DS2L-algorithm with SOM allows automatic cluster number determination and great flexibility in terms of cluster size. It further allows the user to qualitatively determine the level of detail of the clustering result. The application of two SOM-Ensembles helps to remove arbitrariness from the feature selection process, also a common issue in feature-based clustering, as well as to obtain robust and practice-oriented results even for groundwater observation networks that are subject to change over time. The combination of these methods, therefore, creates a solid clustering framework with advantages in terms of (i) making use of heterogeneous data (ii) operating in a comparably high-automated manner, still leaving adaption possibilities to specific dataset characteristics and analysis goals, as well as (iii) obtaining robust, practice-oriented results. The presented framework is easily transferable to other time series-clustering applications in various domains by exchanging the describing features. For cluster ordering and visualization, we propose the use of a weighted correlation measure ($\overline{R_W}$).

The clustering results illustrate the above characteristics well. Similar dynamic patterns are derived from a large data set, which can be used for further processing (e.g., forecasting) and interpretation. Our results also show that the frequently made assumption that nearby wells have a more similar dynamic than wells further apart is only partly true, even for wells in the same aquifer. Moreover, there are similar dynamic patterns in some cases with no clear spatial reference, making it important to cluster wells according to their dynamics rather than spatial proximity or common aquifer properties.

We confirmed that groundwater dynamics are a complicated interaction of most diverse factors, where some of them are hard to determine or are even poorly understood at all. This makes disentangling the contributions usually hard, not to mention the mostly incomplete information on such metadata. We mainly focused on framework development, motivated by the superior goal of selecting representatives for forecasting purposes, which is why it only lies partly within the scope of this work to improve the understanding of the different factors contributing to groundwater dynamics. Thus, we have comparably small or almost no variation in geological conditions, aquifer type, and similar parameters, which is not the best starting point for a search for such correlations. Nevertheless, we hope that our approach can contribute to this general question, besides the improved system knowledge on the local scale, which a hydrograph grouping itself already provides. This applies especially

because studies of groundwater dynamics and their connections to relevant driving forces are comparatively rare yet (Giese et al., 2020). To fully exploit the potential of this method in contributing to the improvement of system knowledge, comprehensive data sets of potential influencing factors covering the complete study area should be available. The goal should be to link driving forces directly to features or indices. For this purpose, more systems should be subject to research studies to explore many different characteristics and system properties. We also imagine that once a better understanding of dynamic-controlling factors is in place, a prediction of ungauged locations generally seems possible.

Acknowledgments

We thank Michel Wingering (LUBW) for insightful assessments and discussions, and Guénaël Cabanes for kindly providing the scripts of his algorithms and the permission to republish them.

Chapter III

Groundwater Level Forecasting with ANNs – A Model Comparison

This chapter is based on a study published in Hydrology and Earth System Sciences (HESS) and is an edited reprint of:

Wunsch, A., Liesch, T., Broda, S., 2021. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrology and Earth System Sciences 25, 1671–1687. doi: 10.5194/hess-25-1671-2021

The following links provide access to the associated online resources of this study:

Paper

Electronic Supplementary Material

Code

1 Introduction

Groundwater is the only possibility for 2.5 billion people worldwide to cover their daily water needs (UNESCO, 2012), and at least half of the global population uses groundwater for drinking water supplies (WWAP, 2015). Moreover, groundwater also constitutes a substantial amount of global irrigation water (FAO, 2010), which altogether and among other factors such as population growth and the climate crisis, make it a vital future challenge to dramatically improve the way of using, managing, and sharing water (WWAP, 2015). Accurate and reliable groundwater level forecasts are a key tool in this context, as they provide important information on the quantitative availability of groundwater and can thus form the basis for management decisions and strategies.

Especially due to the success of DL approaches in recent years and their more and more widespread application in our daily life, DL starts to transform traditional industries and is also increasingly used across multiple scientific disciplines (Shen, 2018a). This applies as well to water sciences, where ML methods, in general, are used in a variety of ways, as data-driven approaches offer the possibility to directly address questions on relationships between relevant input forcings (e.g., precipitation) and important system variables (e.g., runoff, or GWL), without the need to build classical models and explicitly define physical relationships. This is especially handy because these classical models sometimes might be oversimplified or, in the case of numerical models, data-hungry, difficult and time-consuming to set up and maintain, and therefore expensive. In particular, ANNs have been successfully applied to a variety of surface water (Maier et al., 2010) and groundwater level (Rajaee et al., 2019) related research questions already; however, especially DL was used only gradually at first (Shen, 2018a), but is just about to take off, which is reflected in the constantly increasing number of DL and water resource-related publications (see, e.g., Chen et al., 2020; Duan et al., 2020; Fang et al., 2020, 2019; Gauch et al., 2020, 2021; Klotz et al., 2020; Kraft et al., 2020; Kratzert et al., 2019a, 2018, 2019b; Pan et al., 2020; Rahmani et al., 2021). In this work we explore and compare the abilities of (shallow) NARX models, to the currently popular DL approaches LSTM and CNN. During the last years several authors have shown the ability of NARX to successfully model and forecast groundwater levels (Alsumaiei, 2020; Chang et al., 2016; Di Nunno and Granata, 2020; Guzman et al., 2017, 2019; Hasda et al., 2020; Izady et al., 2013; Jeihouni et al., 2019b; Jeong and Park, 2019; Wunsch et al., 2018; Zhang et al., 2019). Although LSTMs and CNNs are state-of-the-art DL techniques and are commonly applied in many disciplines, they are not yet widely adopted in groundwater level prediction applications, if at all, mainly in the last two years. LSTMs were used twice as often to predict groundwater levels (Afzaal et al., 2020; Bowes et al., 2019; Jeong and Park, 2019; Jeong et al., 2020; Müller et al., 2020; Supreetha et al., 2020; Zhang et al.,

2018) compared to CNNs (Afzaal et al., 2020; Lähivaara et al., 2019; Müller et al., 2020). The main reason might be that the strength of CNNs is mainly the extraction of spatial information from image-alike data, whereas LSTMs are especially suited to process sequential data, such as from time series. Overall, these studies show that LSTMs and CNNs are very well capable of forecasting groundwater levels. Both, Afzaal et al. (2020) and Müller et al. (2020) also directly compared the performance of LSTMs and CNNs but no clear superiority of one to the other can be drawn from their results. Müller et al. (2020), who focus on hyperparameter optimization, draw the conclusion that CNN results are less robust compared to LSTM predictions, however, other analyses in their study also show better results of CNNs compared to LSTMs. Jeong and Park (2019) conducted a comparison of NARX and LSTM (among others) performance on groundwater level forecasting. They found both to be the best models in their overall comparison concerning the prediction accuracy, however, they used a deep NARX model with more than one hidden layer. To the best of the authors' knowledge, no direct comparison has yet been made of (shallow) NARX, LSTMs, and CNNs to predict groundwater levels.

This study aims to provide an overview of the predictive ability in groundwater levels of shallow conventional recurrent ANN, namely NARX, and popular state-of-the-art DL-techniques LSTM and (1D-)CNN. We compare the performance on single value (sequence-to-value (seq2val), also known as one-step-ahead, sequence-to-one, or many-to-one) and sequence (sequence-to-sequence, seq2seq) forecasting. We use data from 17 groundwater wells within the Upper Rhine Graben region in Germany and France, selected based on prior knowledge and representing the region's total bandwidth of groundwater dynamic types (Wunsch et al., 2022b, or chapter II). Further, we use only widely available and easy-to-measure meteorological input variables (precipitation, temperature, and relative humidity), making our approach widely applicable and highly transferable. All models are optimized using Bayesian optimization models, which we extend to also solve the common input variable selection problem by considering the inputs as optimizable parameters. Further, the data-dependency of all models is explored in a simple experimental setup for whether there are substantial differences between shallow and deep learning models regarding their need for training data, as one might suspect.

2 Methodology

2.1 Input Variables

In this study, we only use the meteorological input variables precipitation (P), temperature (T), and relative humidity (rH), which in general are widely available and easy to measure. In principle, this makes this approach easily transferable and thus applicable almost everywhere. Precipitation may serve as a surrogate for GW recharge; temperature, and relative humidity include the relationship of GW to evapotranspiration and at the same time provide the network with information on seasonality due to the usually distinct annual cycle. As an additional synthetic input variable, a sinusoidal signal fitted to the temperature curve (Tsin), can provide the model with noise-free information on seasonality, which often allows considerably improved predictions to be made (Kong-A-Siou et al., 2014). Without a doubt, the most important input variable out of these is P, since GW recharge usually has the greatest influence on GW dynamics. Therefore, P is always used as an input variable, the suitability of the remaining variables is checked and optimized for each time series and each model individually. The fundamental idea is that for wells with primarily natural GW dynamics, the relationship between groundwater levels and the important processes of GW recharge and evapotranspiration should be mapped via the meteorological variables P, T, and rH. However, especially for wells with a dynamic influenced by other factors, this is usually only valid to a limited extent since groundwater dynamics can depend on various additional factors such as groundwater extractions or surface water interactions (chapter II). Due to a typically strong autocorrelation of GWL time series, a powerful predictor for the future groundwater level is the groundwater level in the past. Depending on the purpose and methodological setup, it does not always make sense to include this variable; however, where meaningful, we also explored past GWL as inputs (GWL_{t-1}).

2.2 Nonlinear Autoregressive Exogenous Model

NARX models relate the current value of a time series to past values of the same time series as well as to current and past values of additional exogenous time series. We implement this type of model as a recurrent neural network, which extends the well-known feed-forward MLP structure (Figure I.3) by a global feedback connection between output and input layer (Figure III.1). One can therefore also refer to it as recurrent MLP. NARX are frequently applied for nonlinear time series prediction and nonlinear filtering tasks (Beale et al., 2016). Similar to other types of RNNs, NARX have difficulties in capturing long-term dependencies due to the problem of vanishing and exploding gradients (Bengio et al., 1994), yet they

can keep information up to three times longer than simple RNNs (Lin et al., 1996, 1995), so they can converge more quickly and generalize better in comparison (Lin et al., 1998). Using the recurrent connection, future outputs are both regressed on independent inputs and on previous outputs (GWLs in our case), which is the standard configuration for multi-step prediction and also known as closed-loop configuration. However, NARX can also be trained by using the open-loop configuration, where the observed target is presented as an input, instead of feeding back the estimated output. This configuration can make training more accurate and efficient, as well as computationally less expensive because learning algorithms do not have to handle recurrent connections (Moghaddamnia et al., 2009). However, experience shows that both configurations can be adequate for training a NARX model since open-loop training often results in more accurate performance in terms of mean errors. In contrast, closed-loop trained models are often better at capturing a time series's general dynamics. NARX also contain a short-term memory, i.e., delay vectors for each input (ID) (and feedback (FD)) (Figure III.1), which allow the availability of several input time steps simultaneously, depending on the length of the vector. Usually, delays are crucial for the performance of NARX models. Please note that some of our experiments include past GWLs for training (compare section 2.1), which is also performed in closed-loop setup and thus uses both multiple observed past GWLs (according to the size of ID) as an input, as well as multiple simulated GWLs (according to the size of FD) via the feedback connection. In a way, this mimics the open-loop setup, however, we still use the feedback connection and simply treat the past observed GWL as an additional input variable.

The given configuration describes sequence-to-value forecasting; to perform sequence-to-sequence forecasts, some modifications are necessary. Like other ANNs, NARX are capable of performing forecasts of a complete sequence at once, i.e., one output neuron predicts a vector with multiple values. Technically it is necessary to use sequenced inputs with the same length as for the output sequences (here: 12 steps). To build and apply NARX models, we use MATLAB 2020a (Mathworks Inc., 2020) and its Deep Learning Toolbox.

2.3 Long Short-Term Memory

LSTM networks are recurrent neural networks, which are typically applied to model sequential data like time series or natural language (e.g., Chen et al., 2017a). As stated, RNNs suffer from the vanishing gradient problem during backpropagation, and in the case of simple RNNs, their memory barely includes the previous ten time-steps (Bengio et al., 1994). LSTMs, however, can remember long-term dependencies because they have been explicitly designed to overcome this problem (Hochreiter and Schmidhuber, 1997). Besides the hidden state of RNNs, LSTMs contain a cell memory (or cell state) to store information and three gates to

control the information flow (Hochreiter and Schmidhuber, 1997). The forget gate (Gers et al., 2000) controls which and how much information of the cell memory is forgotten, the input gate controls which inputs are used to update the cell memory, and the output gate controls which elements of the cell memory are used to update the hidden state of the LSTM cell. The cell memory enables the LSTM to handle long-term dependencies because information can remain in the memory for many steps (Hochreiter and Schmidhuber, 1997). Several LSTM layers can be stacked on top of each other in a model, however, the last LSTM layer is followed by a traditional fully connected dense layer, which in our case is a single output neuron that outputs the groundwater level. To realize sequence forecasting, as many output neurons in the last dense layer as steps in the sequence are needed. For LSTMs we rely on Python 3.8 (van Rossum, 1995) in combination with the libraries Numpy (van der Walt et al., 2011), Pandas (McKinney, 2010; Reback et al., 2020), scikit-learn (Pedregosa et al., 2011) and Matplotlib (Hunter, 2007). Further we use the Deep-Learning frameworks TensorFlow (Abadi et al., 2015) and Keras (Chollet, 2015).

2.4 Convolutional Neural Networks

CNNs (LeCun et al., 2015) are predominantly used for image recognition and classification (as 2D models) (e.g., Cai et al., 2016; Li et al., 2014). However, in a 1D configuration, they also work well on signal processing tasks, such as natural language processing (e.g., Kiranyaz et al., 2019; Yin et al., 2017). CNNs usually comprise three different layers. Convolutional layers, the first type, consist of filters (or kernels) and feature maps. The input to a filter is called the receptive field and has a fixed size. Each filter (size three in our case) is dragged over the entire previous layer's resulting in an output, which is collected in the feature map. Convolutional layers are often followed by pooling layers that perform down-sampling of the previous layers feature map. Thus, information is consolidated by moving a receptive field over the feature map. Such fields apply simple operations like averaging or maximum selection. Similar to LSTM models, multiple convolutional and pooling layers in varying order can be stacked on top of each other in deeper models. The last layer is followed by a fully connected dense layer with one or several output neurons to match the desired output dimension. To realize sequence forecasting, as many output neurons in the last dense layer as steps in the sequence are needed. For CNNs, we equally to LSTMs use Python 3.8 (van Rossum, 1995) in combination with the libraries and frameworks mentioned above.

2.5 Model Calibration and Evaluation

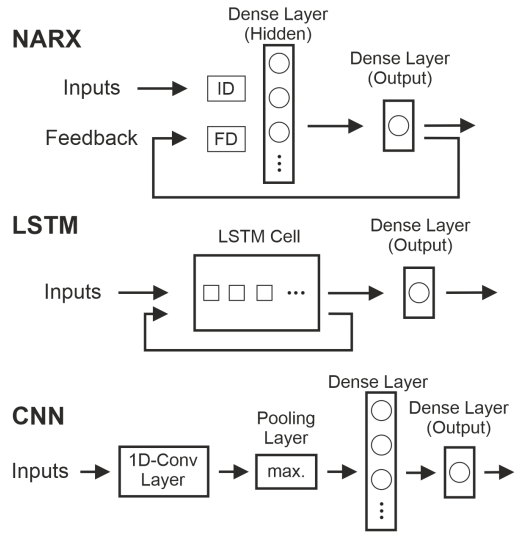
In this study, we use NARX models with one hidden layer and train them in closed-loop using the Levenberg-Marquardt algorithm, which is a fast and reliable second-order local method (Adamowski and Chan, 2011). We choose the closed-loop configuration for training because other hyperparameters (HPs) are optimized using a Bayesian model (see below), which seems to work properly only in a closed-loop configuration, probably due to the artificially pushed training performance in an open-loop configuration. Optimized HPs are the inputs T, Tsin, and rH (1/0, i.e., yes/no), size of the input delays (ID P, ID T, ID Tsin, ID rH), size of the feedback delay vector (FD), and the number of hidden neurons (hidden size). ID and FD can take values between 1 and 52 (which is one year of weekly data), the number of hidden neurons is optimized between 1 and 20. Strictly speaking, input selection is no hyperparameter optimization problem, however, the algorithm can also be applied to select an appropriate set of inputs (Figure III.1). This assumption applies in our study also to LSTM and CNN models.

We choose our LSTM models to consist of one LSTM layer, followed by a fully connected dense layer with a single output neuron in the case of sequence-to-value forecasting. We use Adam-Optimizer with an initial learning rate of 1E-3 and apply gradient clipping to prevent gradients from exploding. Hyperparameters being optimized by a Bayesian model are: the number of units within the LSTM layer (hidden size, 1 to 256), the batch size (1 to 256), and the sequence length (1 to 52). The latter can be interpreted more or less as equivalent to the delay size of the NARX models and is often referred to as the number of inputs (Figure III.1).

The CNN models we apply consist of one convolutional layer, a max-pooling layer, and two dense layers, where the second one consists only of one neuron in the case of sequence-to-value forecasting. Adam-optimizer is used with the same configuration as for the LSTM models. For all CNN models, we use a kernel size of 3 and optimize the batch size (1 to 256), sequence length (1 to 52), the number of filters (1 to 256) within the convolutional layer, as well as the number of neurons within the first dense layer (dense size, 1 to 256) according to a Bayesian optimization model (Figure III.1).

Hyperparameter Optimization is conducted by applying Bayesian optimization using the Python implementation by Nogueira (2014). We apply 50 optimization steps as a minimum (25 random exploration steps followed by 25 Bayesian optimization steps). After that, the optimization stops as soon as no improvement has been recorded during 20 steps or after a maximum of 150 steps. For the NARX models, we use the MATLAB built-in Bayesian-optimization, where the first 50 steps cannot be distinguished as explained above, however, the rest applies accordingly. The acquisition function in all three cases is expected

a)



b)

	optimized parameters	possible values
NARX	Hidden Size	1 to 20
	Input Delay - ID (for each Input)	1 to 52
	Feedback Delay - FD	1 to 52
LSTM	Hidden Size	1 to 256
	Seq.-Length	1 to 52
	Batch Size	1 to 256
CNN	Dense Size	1 to 256
	Seq.-Length	1 to 52
	Batch Size	1 to 256
	No. of filters	1 to 256
all models	T - Input	yes or no
	rH - Input	yes or no
	Tsin - Input	yes or no

Figure III.1: **a)** Simplified schematic summary of the models and their structures used in this work. ID/FD are delays, circles in dense layers symbolize neurons, squares within the LSTM cell the number of hidden units respectively; **b)** Hyperparameters (and inputs) of each model used to tune the models by using Bayesian optimization algorithm, the last column summarizes the optimization ranges for each parameter.

improvement, and the optimization target function we chose is the sum of Nash-Sutcliffe efficiency (NSE) and squared Pearson's correlation coefficient (R^2) (compare equations III.1 and III.2) because these two criteria are important and well-established criteria for assessing the forecast accuracy in water-related contexts.

All three model types use 30 as the maximum number of training epochs. To prevent overfitting, we apply early stopping with a patience of five steps. The testing or evaluation period in this study for all models are the years 2012 to 2015 (inclusively). This period is exclusively used for testing the models. The data before 2012 are of varying length (hydrographs start between 1967 and 1994, see also Figure III.3) and split into three parts, namely 80% for training, and as well 10% for early stopping as 10% for testing during HP-Optimization (opt-set) (Figure III.2). Thus, the target function of the HP-optimization procedure is only calculated on the opt-set.

All data are scaled between -1 and 1, and all models are initialized randomly. They show, therefore, a dependency towards the random number generator seed. To minimize initialization influence, we repeat every optimization step 5 times and take the mean of the target function. For the final model evaluation in the test period (2012–2016), we use ten pseudo-random initializations and calculate errors of the median forecast. For sequence2sequence forecasting, we always take the median performance over all forecasted sequences, which each have a length of 3 months or 12 steps, respectively. This is a realistic length for

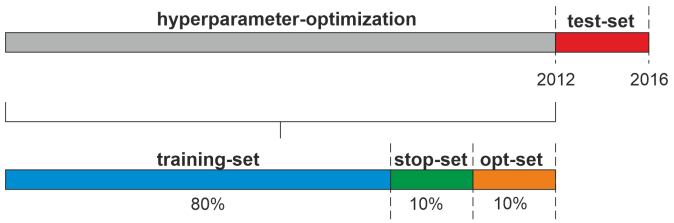


Figure III.2: Data splitting scheme: each time series is split into four parts for training, early stopping, HP optimization as well as testing. The latter is fixed to the period years 2012 to 2016 for all wells, the former three parts depend on the available time series length.

direct sequence forecasting of groundwater levels, which also has some relevance in practice because it (i) provides useful information for many decision-making applications (e.g., groundwater management), and (ii) is also an established time-span in meteorological forecasting, known as seasonal forecasts. In principle, this also allows a performance comparison of 12-step seq2seq forecasts with a potential 12-step seq2val forecast, based on operational meteorological forecasting, where the input uncertainty potentially lowers the groundwater level forecast performance. However, this is beyond the scope of this study, which focuses on neural network architecture comparison.

To judge forecast accuracy, we calculate: Nash-Sutcliffe efficiency, squared Pearson's correlation coefficient, absolute and relative root mean squared error (RMSE/rRMSE), absolute and relative Bias (Bias/rBias) as well as persistency index (PI). For the following equations applies that o represents observed values, as well as p represents predicted values, n stands for the number of samples.

$$NSE = 1 - \frac{\sum_{i=1}^{n} (o_i - p_i)^2}{\sum_{i=1}^{n} (o_i - \bar{o})^2} \tag{III.1}$$

Please note that we use the mean observed values in the denominator until the start of the test period (2012 in the case of our final model evaluation). This represents best the meaning of the NSE, which compares the model performance to the mean values of all known values at the time of the start of the forecast.

$$R^2 = \left(\frac{\sum_{i=1}^{n} (o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n} (o_i - \bar{o})^2} \sqrt{\sum_{i=1}^{n} (p_i - \bar{p})^2}} \right)^2 = r^2 \tag{III.2}$$

In our case, we use the squared Pearson correlation coefficient r^2 as a general coefficient of determination R^2, since it compares the linear relation between simulated and observed GWLs.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \tag{III.3}$$

$$rRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{o_i - p_i}{o_{max} - o_{min}}\right)^2} \tag{III.4}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(o_i - p_i) \tag{III.5}$$

$$rBias = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{o_i - p_i}{o_{max} - o_{min}}\right) \tag{III.6}$$

$$PI = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - o_{last})^2} \tag{III.7}$$

Please note that RMSE and Bias are useful to compare performances for a specific time series among different models, however, only rRMSE and rBias are meaningful to compare model performance between different time series. The persistency index PI basically compares the performance to a naïve model that uses the last known observed groundwater level at the time the prediction starts. This is particularly important to judge the performance when past groundwater levels (GWL_{t-1}) are used as inputs because, especially in this case, the model should outperform a naïve forecast ($PI > 0$).

2.6 Data-dependency

Data-dependency of empirical models is a classical research question (Jakeman and Hornberger, 1993), often focusing on the number of variables but also concerning the length of available data records. Data scarcity is also an important topic in ML in general, especially in DL and the focus of recent research (e.g., Gauch et al., 2021). Therefore, one can expect to find performance differences between both shallow and deep models used in this study. We hence performed experiments to explore the need for training data for each of the model types. For this, we started with a reduced training record length of only two years before testing the performance on the fixed test set of four years (2012-2016). In the following, we gradually lengthened the training record until the maximum available length for each well and tracked the error measure changes. This experiment aims to give an impression of how much data might be needed to achieve satisfying forecasting performance and if there are substantial differences between the models; however, it is out of the scope to answer this very complex question in a general way for each of the modeling approaches.

2.7 Computational Aspects

We used different computational setups to build and apply the three model types. We built the NARX models in MATLAB and performed the calculations on the CPU (AMD-Ryzen 9 3900X). Using a GPU instead of a CPU is not possible for NARX models in our case because of the Levenberg-Marquardt training algorithm, which is not suitable for GPU computation. However, both LSTMs and CNNs can be calculated on a GPU, which in the case of LSTMs is the preferred option. For CNNs, we observed a substantially faster calculation (factor 2 to 3) on the CPU and therefore favored this option. Both LSTMs and CNNs were built and applied using Python 3.8; the GPU we used for LSTMs was a Nvidia GeForce RTX 2070 Super.

3 Data and Study Area

In this study, we examine the groundwater level forecasting performance at 17 groundwater wells within the Upper Rhine Graben area (Figure III.3), the largest groundwater resource in central Europe (LUBW, 2006). The aquifers of the URG cover 80% of the drinking water demand of the region as well as the demand for agricultural irrigation and industrial purposes (Région Alsace - Strasbourg, 1999). The wells are selected from a larger dataset from the region with more than 1800 hydrographs. Based on the analyses of Wunsch et al. (2022b) (chapter II), the wells of this study represent the total bandwidth of groundwater dynamics occurring in the dataset. The whole dataset mainly consists of shallow wells from the uppermost aquifer within the Quaternary sand/gravel sediments of the URG. Mean GWL depths are smaller than 5 m bgl for 70% of the data, rising to a maximum of about 20-30 m towards the Graben edges. The considered aquifers generally show high storage coefficients and high hydraulic conductivities in the order of 10E-4 to 10E-3 m/s (LUBW, 2006). Strong anthropogenic influences exist in some areas, e.g., the northern URG, due to intensive groundwater abstractions and management efforts. A list of all examined wells with additional information on identifiers and coordinates can be found in the supplement (Table S1). All groundwater data are available for free via the web services of the local authorities (HLNUG, 2019; LUBW, 2018; MUEEF, 2018). The shortest modeled time series starts in 1994, the longest in 1967, however, most hydrographs (12) start between 1980 and 1983 (Figure III.3). Meteorological input data was derived from the HYRAS dataset (Frick et al., 2014; Rauthe et al., 2013), which can be obtained free of charge for non-commercial purposes on request from the German Meteorological Service. This study exclusively considers weekly time steps for both groundwater and meteorological data.

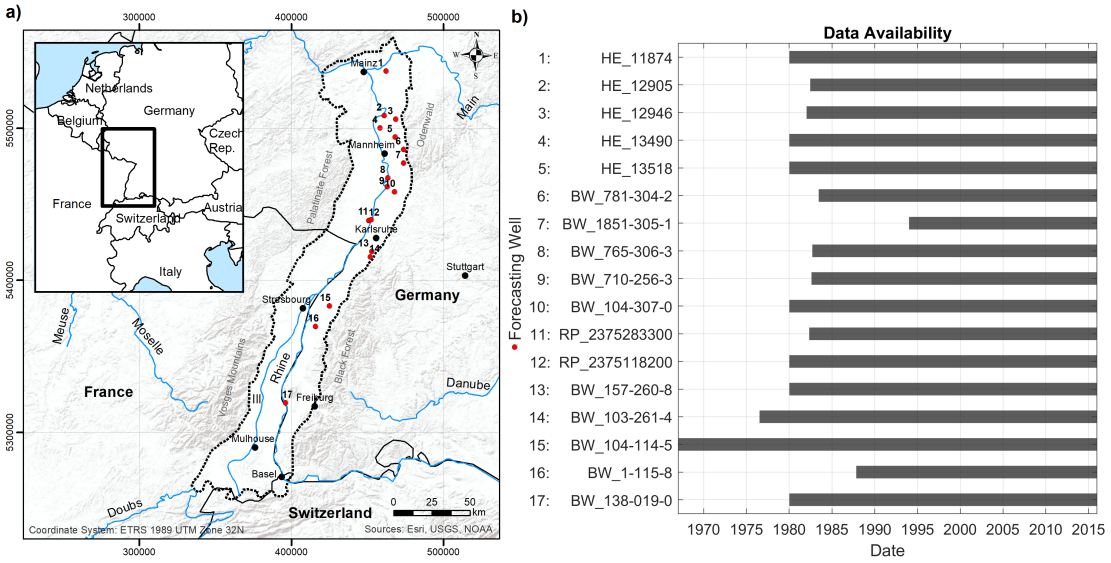


Figure III.3: **a)** Study area and position of examined wells; **b)** respective time series length for each of the wells.

4 Results and Discussion

4.1 Sequence-to-Value Forecasting Performance

Figure III.4 summarizes and compares the overall seq2val forecasting accuracy of the three model types for all 17 wells. Figure III.4a shows the performance when only meteorological inputs are used, the models in Figure III.4b are additionally provided with GWL_{t-1} as an input. Because the GWL of the last step has to be known, the latter configuration has only limited value for most applications since only one-step-ahead forecasts are possible in a real-world scenario. However, the meteorological inputs of the former configuration are usually available as forecasts themselves for different time horizons. Figure III.4a shows that on average, NARX models perform best, followed by CNN models, LSTMs achieve the least accurate results. This is consistent for all error measures except rBias, where CNN models show slightly less bias than NARX. However, all models suffer from considerable negative bias values in the same order of magnitude, meaning that GWLs are systematically underestimated. Providing information about past groundwater levels up to t-1 (GWL_{t-1}) improves the performance of all three models considerably (Figure III.4b). Additionally, performance differences between the models vanish and remain only visible as slight tendencies. This is not surprising, as the past groundwater level is usually a good or even the best predictor of the future GWL, at least for one step ahead, and all models are able to use this information. The general superiority of NARX in the case of Figure III.4a is therefore not totally surprising, as a

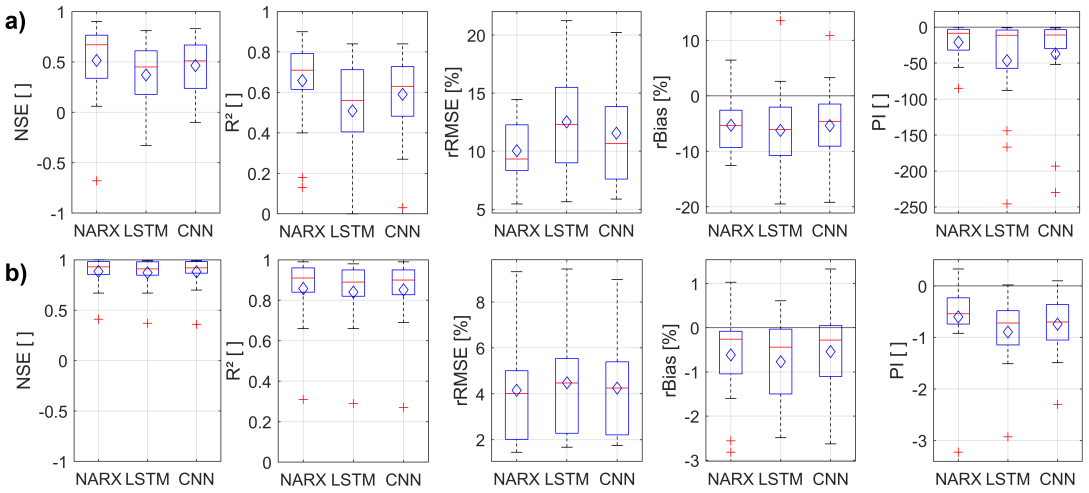


Figure III.4: Boxplots showing the seq2val forecast accuracy of NARX, LSTM and CNN models within the test period (2012-2016) for all considered 17 hydrographs. The diamond indicates the arithmetic mean; **a)** only meteorological inputs; **b)** GWL_{t-1} as additional Input.

feedback connection within the model already provides information of past groundwater levels, even though it also includes a certain forecasting error. However, providing GWL_{t-1} as input to a seq2val-model (Figure III.4b) basically means providing the naïve model itself, which needs to be outperformed in the case of PI metric (compare section 2.5). PI values below zero, therefore, mean that the output is less accurate than the input, which is, apart from the limited benefit for real-world applications mentioned above, why we refrain from further discussion of the models shown in Figure III.4b.

For our analysis, we did not make a pre-selection of hydrographs that show predominantly natural groundwater dynamics and thus a comparatively strong relationship between the available input data and the groundwater level. Therefore, even though hydrographs possibly influenced by additional factors were examined, we can conclude that the forecasting approach in general works quite well, and we reach, e.g., median NSE values of $\geq$0.5 for NARX and CNNs, LSTMs show a median value only slightly lower. In terms of robustness against the initialization dependency of all models (ensemble variability), we clearly observe the highest dependency for NARX, followed by CNN and LSTM, while LSTMs on median perform slightly more robust than CNNs. Including GWL_{t-1} lowers the error variance of the ensemble members, which we used to judge robustness in this case by several orders of magnitude for all models. NARX and LSTMs on median now show slightly lower ensemble variability than CNNs, however, all models are quite close. A corresponding figure was added to the supplement (Figure S69). Furthermore, we added information on the results of the hyperparameter-optimization (Tables S2-S4), a table with all error measure values of each considered hydrograph and model (Table S5), as well as according seq2val forecasting plots (Figures S1 to S34) to the supplement, as well.

Figure III.5 shows exemplarily the forecasting performance of all three models for well BW_104-114-5, where all three models consistently achieved good results in terms of accuracy. The NARX model (a) outperforms both LSTM (b) and CNN (c) models and shows very high NSE and R^2 values between 0.8 and 0.9. The CNN model provides the second-best forecast, which even very slightly shows less underestimation (Bias/rBias) of the GWLs than the NARX model. By comparing the graphs in (a) and (c) we assume that this is only true on average. The CNN model overestimates in 2012 and constantly underestimates the last third of the test period. The NARX model, however, is more consistent and therefore better. Concerning R^2 values, the LSTM basically keeps up with the CNN; all other error measures show the still good, but in comparison worst values. We notice that in accordance with our overall findings mentioned above, the LSTM shows the lowest ensemble variability and, therefore, the smallest initialization dependency. Looking at the selected inputs and hyperparameters, we notice that rH does not seem to provide important information and was therefore never selected as an input. Further, the input sequence length of both LSTM and CNN is equally 35 steps (weeks). In the NARX model, there is no direct correspondence, but a similar value is shown by the parameter FD, and thus the number of past predicted GWL values available via the feedback connection.

In contrast to the above-mentioned well, hardly any systematic can be derived from the choice of input variables across all wells that even might have physical implications for each site. Instead, it is noticeable that certain model types seem to prefer also certain inputs. For example, temperature is only selected as input in 5 out of 17 cases for LSTM models, and in 2 out of 17 cases for CNN models. Furthermore, relative humidity (rH) is always selected for LSTM models except for two times. In the case of NARX models, there seems to be a lack of systematic behavior. For more details, please see Tables S2-S4 in the supporting material.

Our approach assumes a groundwater dynamic mainly dominated by meteorological factors. We can assume that all three model types are basically capable of modeling groundwater levels very accurately if all relevant input data can be identified. To exemplarily show the influence of additional input variables, which, however, are usually not available as input for a forecast or even have insufficient historical data, Figure III.6 illustrates the considerably improved performance after including the Rhine River water level (W), a large streamflow within the study area, using the example of the NARX model for well BW_710-256-3, which indeed is located close to the river. Besides improved performance, we also observe lower variability of the ensemble member results, thus, lower dependency to the model initialization, which also corresponds to other time series, where we often find smaller influence the more relevant the input data are. This also confirms that little accuracy is presumably due to insufficient input data on a case-by-case basis, not necessarily because of an inadequate modeling approach. Similarly, this applies also to other wells in our dataset that show unsat-

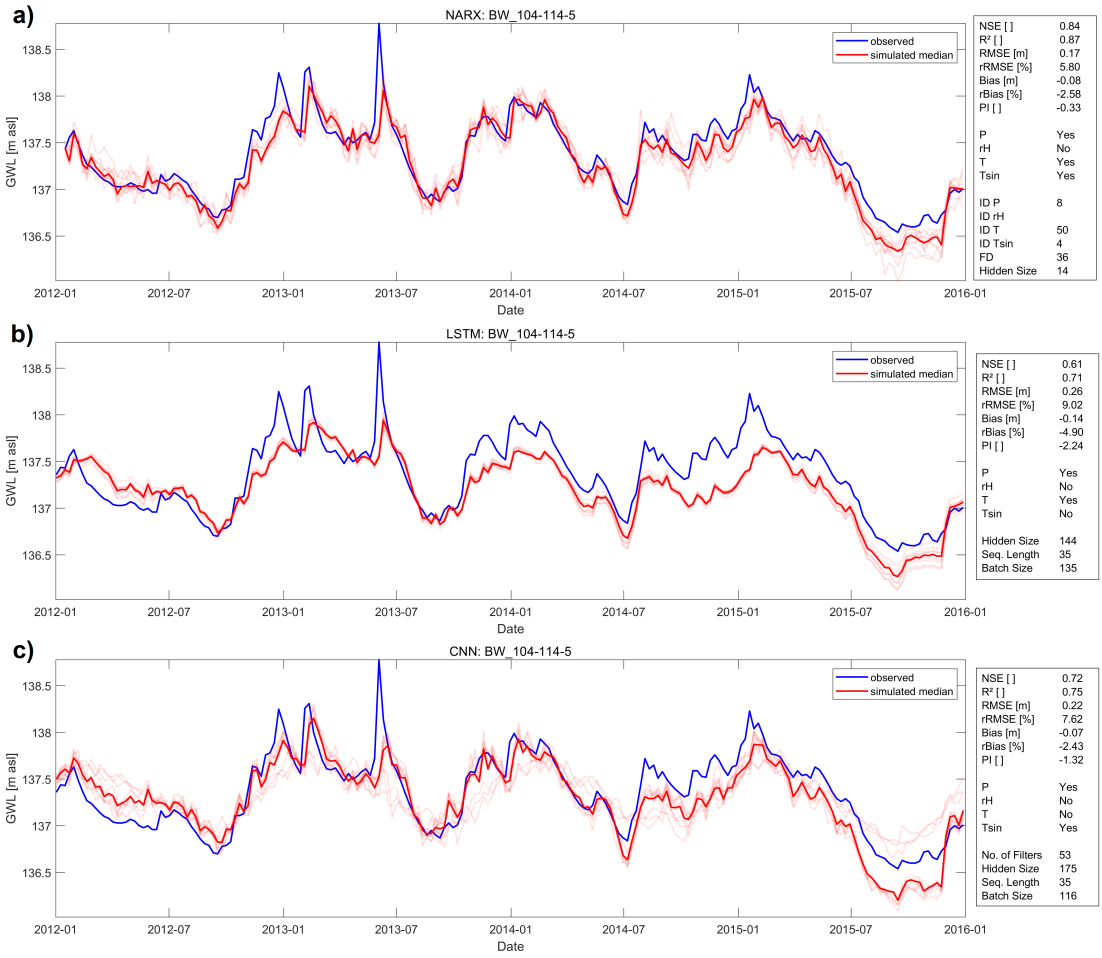


Figure III.5: Forecasts of **a)** a NARX, **b)** a LSTM and **c)** a CNN model for well BW_104-114-5 during the test period 2012-2016.

isfying forecasting performance. Examples of this are wells in the northern part of the URG (e.g., most wells with the prefix HE), for which our approach is generally more challenging due to strong GW extraction activities in this area, and well BW_138-019-0, which is close to the Rhine River and presumably under the influence of a large ship lock nearby. Additionally, this well is within a flood retention area that is spatially coupled to the ship lock.

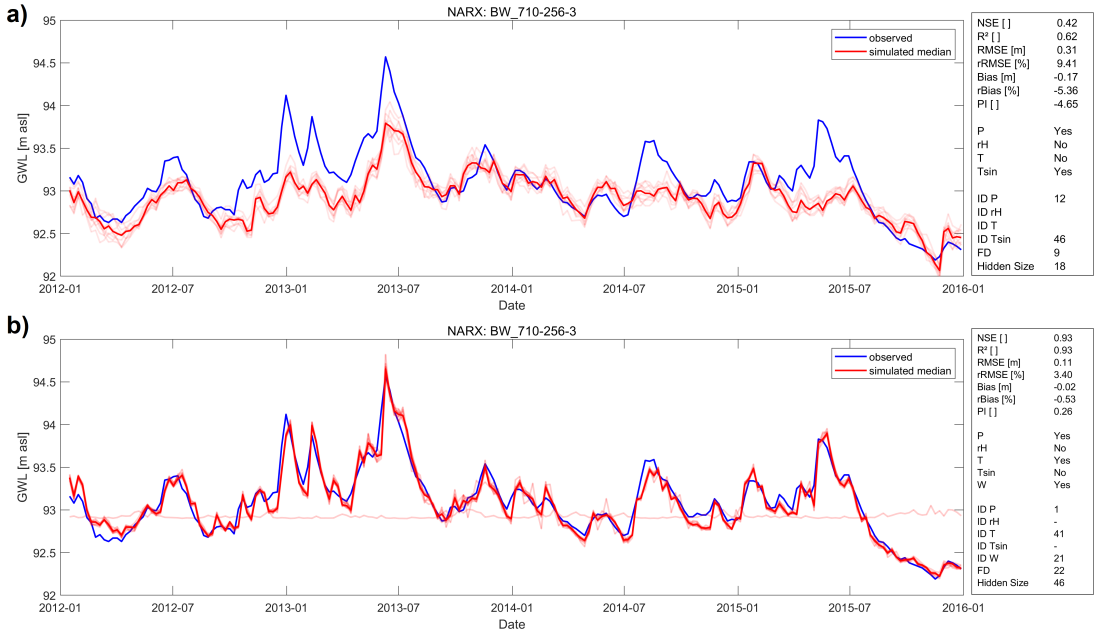


Figure III.6: Forecasting performance exemplarily shown for NARX model of well BW_710-256-3 **a)** based on meteorological input variables and **b)** improved performance after including Rhine River water level (W) as input variable.

4.2 Sequence-to-Sequence Forecasting Performance

Seq2Seq forecasting is especially interesting for short- and mid-term forecasts because the inputs only have to be available until the start of the forecast. Figure III.7 summarizes and compares the overall seq2seq forecasting accuracy of the three model types for all wells. Figure III.7a shows the performance for purely meteorological inputs, Figure III.7b shows the results with an additional GWL_{t-1} input. Equally to the seq2val forecasts (Figure III.4), past GWLs seem to be especially important for LSTM and CNN models as GWL_{t-1} causes substantially improved performance. Without GWL_{t-1}, NARX are superior, presumably due to their inherent global feedback connection. However, NARX show almost equal performance values in both scenarios (Figure III.7a and b). In contrast to seq2val results, for seq2seq forecasts NARX systematically show lower R^2 values than LSTM and CNN models. For all other error measures, the accuracy of NARX models outperforms LSTMs and CNNs in direct comparison for the vast majority of all time series. While LSTMs and CNNs show lower performance for seq2seq forecasting compared to seq2val forecasting, NARX seq2seq models even outperform NARX seq2val models (except for R^2). This is quite counter-intuitive as one would expect it to be more difficult to forecast a whole sequence than a single value. All in all, the scenario including past GWLs (Fig. III.7b) seems to be the preferable one for all three models and shows promising results for real-world applications. Detailed results on all seq2seq models can be found in supplementary Table S6, and Fig. S35-S68.

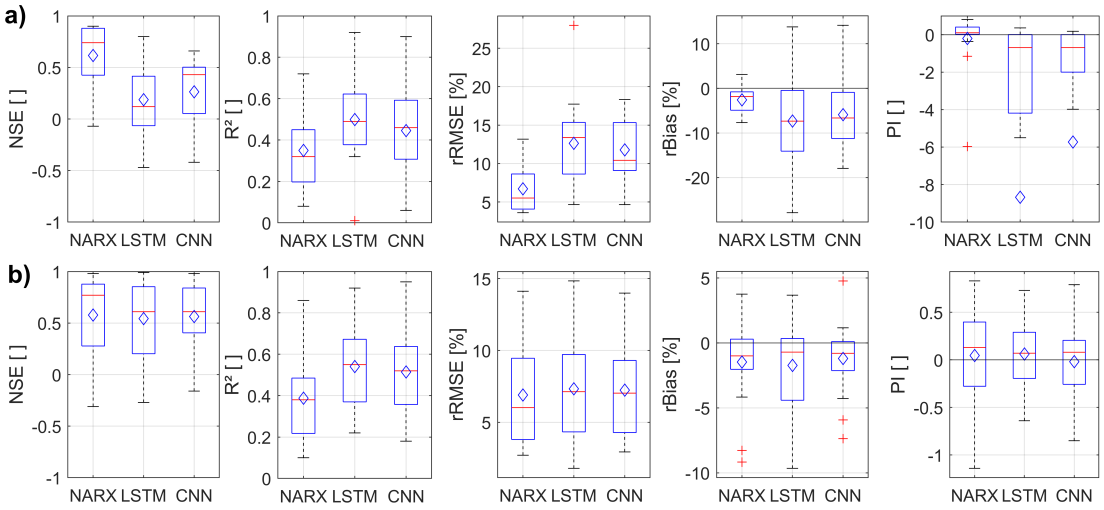


Figure III.7: Boxplots showing the seq2seq forecast accuracy of NARX, LSTM and CNN models within the test period (2012-2016) for all considered 17 hydrographs. The diamond indicates the arithmetic mean; **a)** only meteorological inputs; **b)** GWL_{t-1} as additional Input.

Figure III.8 summarizes exemplarily for well HE_11874 the seq2seq forecasting performance for NARX (a,b), LSTMs (c,d), CNNs (e,f), only with meteorological input variables (a,c,e) and with an additional past GWL input (b,d,f). These confirm that GWL_{t-1} substantially improves the performance of LSTMs and CNNs, however, NARX forecasts in this case only improve very slightly. Especially for LSTMs and CNNs, it is visible that the sequence forecasts of the better models (d,f) mostly estimate the intensity of a future groundwater level change too conservatively; meaning that both in- and decreases are predicted too weak. This is a commonly known issue with ANNs, as extreme values are typically under-represented in the distribution of the training data (e.g Sudheer et al., 2003). We further notice that the robustness of LSTMs and CNNs in terms of initialization dependency, thus the ensemble variability, considerably improves when past GWLs are provided as inputs (Figure III.8). This is also supported by analyzing the ensemble-member error variances and is also true for all other time series in the dataset as well (Figure S69 in the supplement). Just like for seq2val forecasts, NARX usually show considerably lower robustness in terms of initialization dependency; however, the median ensemble performance nevertheless is of high accuracy. Therefore, all models, but especially NARX models, should not be evaluated without including an initialization ensemble. The initialization dependency of LSTMs and CNNs is considerably lower, with LSTMs being even more robust than CNNs.

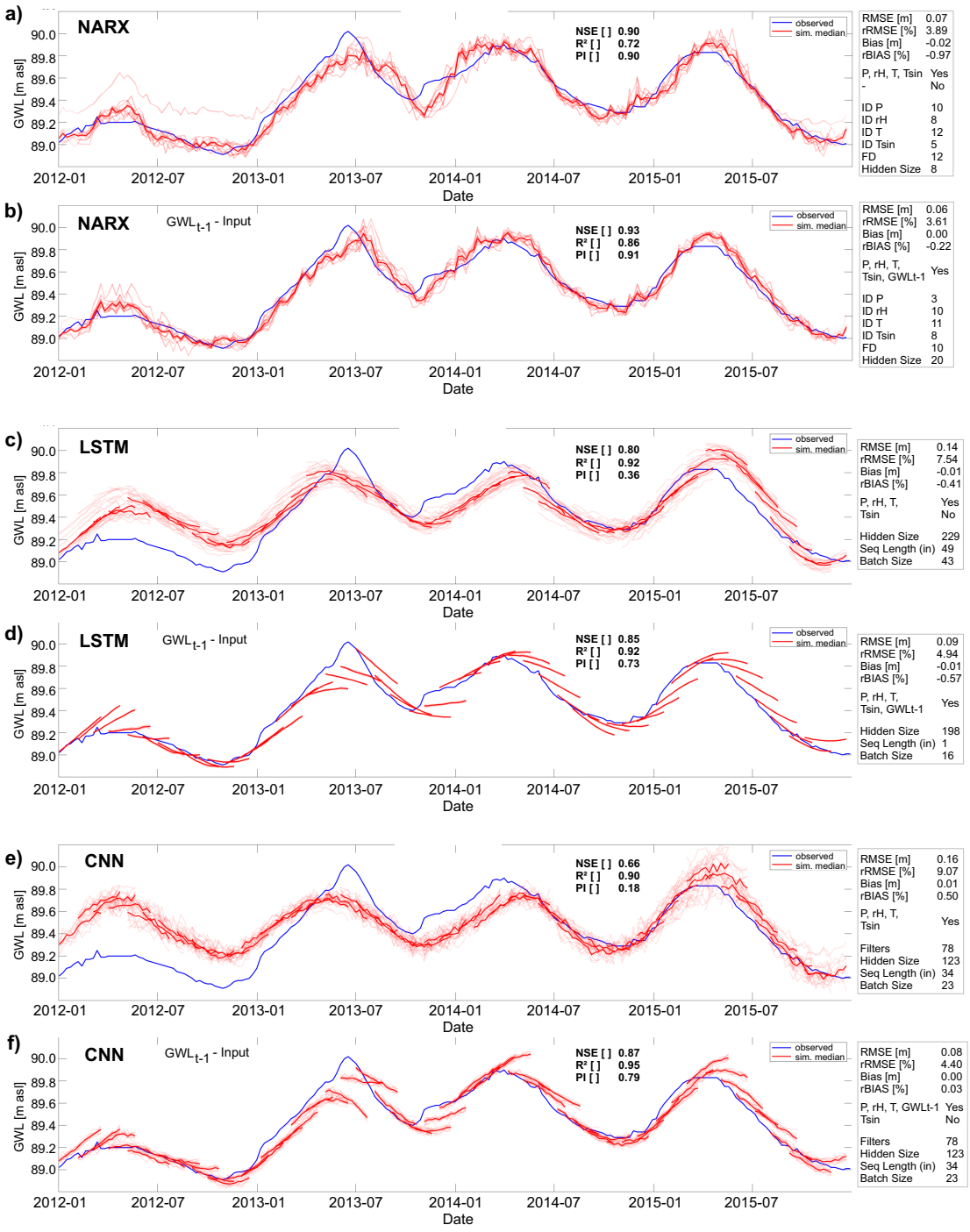


Figure III.8: Forecasts of **a,b)** a NARX, **c,d)** a LSTM and **e,f)** a CNN model for well HE_11874 during the test period 2012-2016; Models in **a,c,e)** use only meteorological input variables, models **b,d,f)** use also past GWL observations

The extraordinary performance of the NARX models, especially in the case of Well HE_11874 (Figure III.8) surprises, because the performance (NSE≥0.9 in both seq2seq models) substantially outperforms the seq2val NARX without GWL_{t-1} input (NSE: 0.35, R^2: 0.75); however, the seq2val NARX model with GWL_{t-1} inputs also showed high accuracy (NSE: 0.99, R^2: 0.99). It is also interesting to note that the sequence predictions of the NARX models overlap exactly and the individual sequences are therefore no longer visible. One reason for this different behavior compared to the LSTM and CNN models is presumably that the technical approach for seq2seq forecasting differs for these models. While LSTMs and CNNs use multiple output neurons to predict multiple time steps, this approach for us did not yield meaningful results for a NARX model, presumably because of feedback connection issues. Instead, we used one NARX output neuron to predict a multi-element vector at once.

4.3 Hyperparameter Optimization and Computational Aspects

During the HP-Optimization, depending on the forecasting approach (seq2val/seq2seq) and available inputs (with/without GWL_{t-1}), there were noticeable differences with regard to the number of iterations required and the associated time needed (Figure III.9). The best parameter combination, especially for CNN and LSTM networks, was often found in 33 steps or fewer, hence after 25 obligatory random exploration steps in only 8 Bayesian steps. Please note that we chose to perform at least 50 optimization steps prior to the analysis, which explains the distribution in the 'total iterations' column. In column two ('best iteration'), we can observe similar behavior of CNNs and LSTMs; NARX are always somehow different from these two. We suspect that this is rather an influence of the software or the optimization algorithm since especially model types implemented in Python show an identical behavior. However, in the majority of cases, the best iteration was found in less than 33 steps, the minimum as well as the maximum number of iteration steps were therefore sufficient. It is interesting that for CNNs and LSTM, the number of steps is similar throughout the experiments, whereas for NARX, the inclusion of GWL_{t-1} as input caused an increase of iterations. Columns three to five in Figure III.9 show substantial differences concerning the calculation speed of the three model types. CNNs outperform all other models systematically, however, concerning the sequence-2-sequence forecasts, NARX models can almost keep up. We also observe that LSTMs seem to slow down when including GWL_{t-1} as input or when performing seq2seq forecasts, the opposite happens in the case of NARX models, which speed up in these cases. This also means that even though NARX models need more optimization iterations until the assumed optimum than LSTMs, in terms of the time, they outperform them due to shorter duration per iteration (col. 3). Please note that it is out of the scope of this work to provide detailed assessments of the calculation speed under benchmark conditions but to share practice-relevant insights for fellow hydrogeologists.

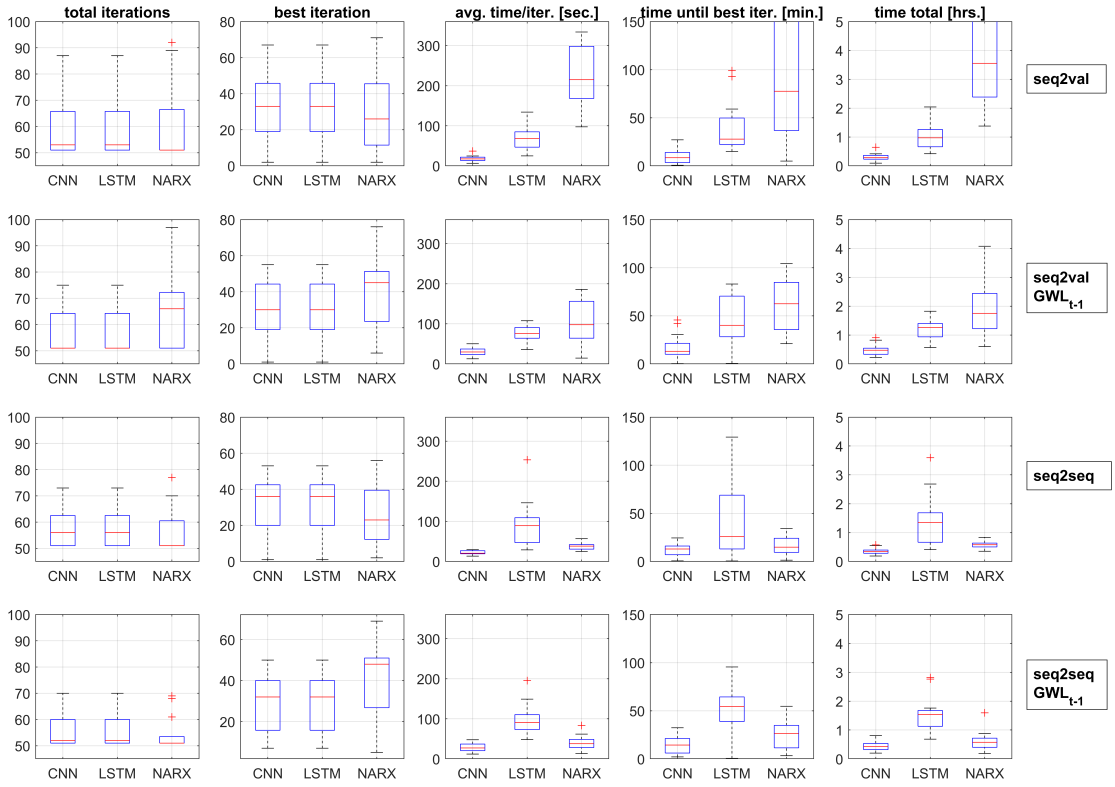


Figure III.9: Comparison of the performed HP-optimizations (columns 1 and 2), their calculation time per iteration in seconds (col. 3), until the optimum was found (minutes) (col. 4) and the total time spent on optimization in hours (col. 5).

4.4 Influence of Training Data Length

In the following section, we explore similarities and differences of NARX, LSTMs, and CNNs in terms of the influence of training data length. It is commonly known that data-driven approaches profit from additional data, however, it still remains an open question how much data are necessary to build models, which are able to perform reasonable calculations. This is because the answer is highly dependent on the application case, data properties (distribution e.g.), and model properties, as model-depth can sometimes exponentially decrease the need for training data (Goodfellow et al., 2016). Therefore, this question cannot be entirely answered by a simple analysis as we perform it here. Nevertheless, we still want to give an impression on how much data might be approximately needed in the case of groundwater level data in porous aquifers and if the models substantially differ in their need for training data. For our analysis, we always consider the forecasting accuracy during the 4-year testing period (2012-2016) and systematically expand the training data basis year by year, starting in 2010, thus with only clearly insufficient two years of training data. We focus on sequence-to-value forecasting due to the more straightforward interpretability of the results, and we

always consider the median performance of 10 different model initializations for evaluation. Figure III.10 summarizes the performance and the improvement that comes with additional training data, all values are normalized per well to make them comparable. Please note that all models at least show 28 years of training data (until 1982), only three models exceed 30 years of training data (1980), thus, the number of samples represented by the boxplots decreases considerably after 30 years. Figure III.10 summarizes as well models with as without GWL_{t-1} inputs because no considerably different behavior was observed for each group. Please find according figures for each group in the supplement (Figures S70-S71).

As expected, we observe considerable improvements with additional training data. NARX models seem to improve more or less continuously and also work better with little data, whereas for LSTMs and CNNs, some kind of threshold is visible (about ten years, thus approx. 500 samples), where the performance considerably increases and rapidly approaches the optimum. It should be noted, though, that this can probably not be transferred to other time steps, i.e., in the case of daily values e.g., 500 days will most certainly not be enough, since only one entire annual cycle is included. We explored the reason for this threshold and observed that when stopping the training five years earlier (2007), the threshold now occurs correspondingly five years earlier (Figure S72 in supplement). Additionally, we found that several standard statistic values such as mean, median, variance, overall maximum, and the 25^{th}, 75^{th}, as well as the 97.5^{th} quantile show similar thresholds (Figure S73 in supplement). Thus, the early years of the 2000s seem to be especially relevant for our test period. This is a highly dataset-specific observation that cannot be generalized; however, this also shows that it is vital to include relevant training data, which is, however, not very easy to identify. Nevertheless, as a rule of thumb, the chance of using the correct data, increases with the amount of available data. These findings are supported by the observation that not every additional year improves the accuracy, only the overall trend is positive. This seems plausible because, especially when conditions change over time, the models can also learn behavior that is no longer valid, possibly decreasing future forecast performance. One should, therefore, not only include as much data as possible but also carefully evaluate and also possibly shorten the training data basis necessary.

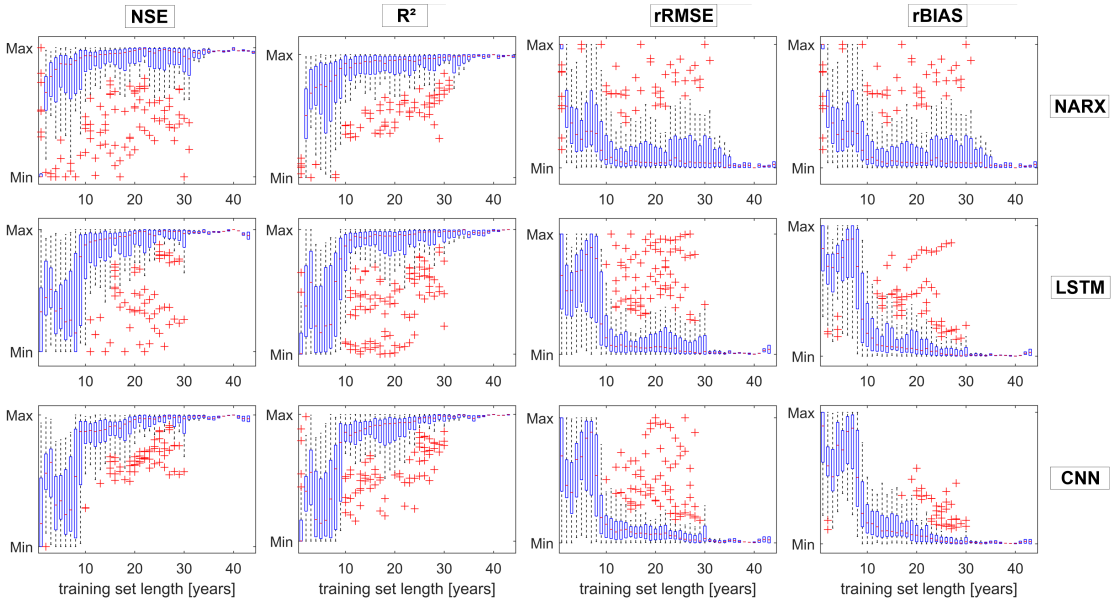


Figure III.10: Influence of training data length on model performance.

5 Conclusions

In this study, we evaluate and compare the groundwater level forecasting accuracy of NARX, CNN, and LSTM models. We examine as well sequence-to-value as sequence-to-sequence forecasting scenarios. We can conclude that in the case of seq2val forecasts, all models are able to produce satisfying results, and NARX models, on average, perform best, LSTMs the worst. Since CNNs are much faster in calculation speed than NARX and only slightly behind in terms of accuracy, they might be the favorable option if time is an issue. If accuracy is especially important, one should stick with NARX models. LSTMs, however, are most robust against initialization effects, especially compared to NARX. Including past groundwater levels as inputs strongly improves CNN and LSTM seq2val forecast accuracy. However, all three models mostly cannot beat the naïve model in this scenario and are therefore of no value.

Especially when no input data are available in short- and mid-term forecasting applications, sequence-to-sequence forecasting is of particular interest. Again, past groundwater levels as input considerably improved CNN and LSTM performance, NARX performed almost similar in both scenarios. Overall, NARX models show the best performance (except R^2 values) in the vast majority of all cases. In addition to the fast calculation of NARX in this case, which almost keeps up with CNN speed, they are clearly preferable. However, NARX models are least robust against initialization effects, which nevertheless are easy to handle by implementing a forecasting ensemble.

We further analyzed what data basis might be needed or sufficient to reach acceptable results. As expected, we found that in principle, the longer the training data, the better; however, a noteworthy threshold seems to exist for about ten years of weekly training data, below which the performance becomes considerably worse. This applies especially for LSTM and CNN models but was also found to presumably be highly dataset specific. Overall, NARX seem to perform better in comparison to CNN and LSTM models when only little training data are available.

The results are surprising in a way that LSTMs are widely known to perform especially well on sequential data and are therefore also more commonly applied. In this work, they were outperformed by CNNs and NARX models. We showed that for this specific application (i) CNNs might be the better choice due to considerably faster calculation and mostly similar performance, and (ii) even though DL-approaches are currently often preferred over traditional (shallow) neural networks such as NARX, the latter should not be neglected in the selection processes especially when there is little training data available. Particularly NARX sequence-to-sequence forecasting seems to be promising for short- and mid-term forecasts. However, we do not want to ignore the fact that LSTMs and CNNs might perform substantially better with a larger dataset, which better fulfills common definitions of DL-applications and where deeper networks can demonstrate their strengths, such as automated feature extraction. Since such data are usually not available in groundwater level prediction tasks yet, for the moment, this remains in theory.

Chapter IV

Groundwater in the Context of Climate Change

The following chapter is based on a study published in Nature Communications and is an edited reprint of:

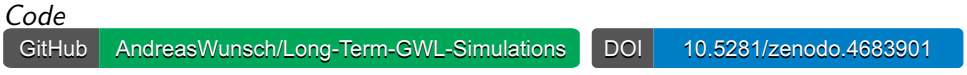
Wunsch, A., Liesch, T., Broda, S., 2022. Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. Nature Communications, 13, 1221. doi: 10.1038/s41467-022-28770-2

The following links provide access to the associated online resources of this study:

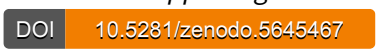
Paper

Electronic Supplementary Material

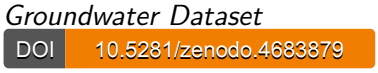
Code



Additional Supporting Information



Groundwater Dataset



1 Introduction

The climate crisis is increasingly altering water availability even in generally water-rich areas like Germany, where overall water stress is currently low (UBA, 2020). Nevertheless, hot and dry summers in recent years (especially 2018-2020) led to ongoing exceptional droughts (UFZ, 2021; Wriedt, 2020) with severe consequences for agriculture and ecology, such as drought damages in forests, reduced crop yields, and extreme low flows in rivers. Drought effects accumulated over the years because winter precipitation did not compensate for summer deficits. This applies not only but especially to groundwater resources, which constitute the major source of drinking water supply in Germany (almost 70%) (Destatis, 2021). Declining groundwater levels due to generally reduced groundwater recharge and higher water demand in summer regionally forced water suppliers to exploit their current maximum capacity during dry periods to meet the demand; locally, even water supply shortages occurred. During future dry periods, strong usage conflicts can be expected in areas of low water availability between water suppliers and industry (process and cooling water), additionally amplified by increasing agricultural irrigation demand, which currently has only minor significance with less than 2% of the total withdrawal volume (UBA, 2020). Knowledge of future groundwater level development, especially in the long-term, is, therefore, crucial to develop sustainable groundwater management plans to meet future demands, solve usage conflicts and protect ecosystems.

Changing climate affects groundwater in several direct and indirect ways (Taylor et al., 2012). Major direct drivers are changes in precipitation, snowmelt, and evapotranspiration (Wu et al., 2020). Different representative concentration pathways (RCP) describe possible future climate scenarios. The current situation best matches RCP8.5, often described as a business-as-usual scenario with increasing greenhouse gas (GHG) emissions. Despite existing mitigation efforts, this scenario might be the most plausible one for the near future (Schwalm et al., 2020). RCP2.6, a stringent mitigation scenario with an average global warming below 2°C above pre-industrial temperatures (IPCC, 2014), might be hard to reach at all, and even the intermediate RCP4.5 is still more ambitious than current (as of 2021) nationally determined contributions under the Paris Agreement, according to UN-FCCC (UNFCCC, 2021). Their analyses estimate a global warming of approximately 2.7°C compared to pre-industrial temperatures. For Germany, analyses based on climate projections show opposing trends in terms of water availability. With some differences between drier and wetter models, they find a slight increase in annual precipitation sums, i.e., generally more water, but at the same time with high agreement between models a significant temperature increase of several degrees Celsius by 2100 (Jacob et al., 2014; Marx et al., 2017; Thober et al., 2018), i.e., less water. The resulting effect on groundwater resources is therefore not directly clear

and needs to be analyzed. Higher precipitation is generally expected during winter, which in combination with a decreasing amount of snow, thus increasing direct infiltration, leads to higher groundwater recharge during winter and less in spring. For the few snow-dominated regions in Germany (e.g., in the South), this might cause changes in seasonality (Wu et al., 2020), however, overall this plays a minor role. Weather extremes are expected to intensify; therefore, longer droughts and more frequent intense rainfall events will occur (Taylor et al., 2012). Generally, higher temperatures cause higher atmospheric water demand, thus increasing evapotranspiration, which leads to less infiltration and, therefore, less groundwater recharge. Especially unconfined, shallow aquifers are most likely to be sensitive to direct climate change effects (Kløve et al., 2014). Indirect climate change influences on groundwater are mostly related to anthropogenic groundwater withdrawals or associated with land-use changes (Taylor et al., 2012), and it is known that the groundwater storage reduction caused by pumping could easily far exceed natural recharge (de Graaf et al., 2019; Wu et al., 2020). The impact of these factors will be exacerbated as water demand increases to as well meet the needs of regionally growing populations (mainly due to growing urban areas), as of industry and agricultural irrigation. To date, there are no reliable data available that estimate the future development of such factors under different climate change scenarios.

In recent years, ANN approaches have proven their usefulness in predicting groundwater levels (Guzman et al., 2017; Jeong and Park, 2019; Jeong et al., 2020; Müller et al., 2020; Wunsch et al., 2021; Zhang et al., 2020), even using a highly transferable approach with purely climatic input variables (e.g., Wunsch et al., 2021). In a previous study (Wunsch et al., 2021), we showed that 1D-CNNs are a good choice for groundwater level simulation because they mostly outperform even LSTM models in terms of accuracy and calculation speed, as well as they showed considerably higher stability, flexibility, and calculation speed compared to NARX models. Therefore, they are an accurate, fast, and reliable method of choice for this study. Unlike physically-based models, which usually require a very good knowledge of local conditions and need to be time-consumingly built and calibrated, data-driven models such as ANNs can predict a target variable using only relevant driving forces. This makes studies on larger areas easier and is, therefore, the favored approach for this study. To the authors' knowledge, no comprehensive direct evaluation of groundwater level development until 2100 exists for Germany yet. Besides a rather old small-scale study (Eckhardt and Ulbrich, 2003) also a regional-scale study for the Danube basin has been conducted to date (Barthel et al., 2012). The latter uses several dynamically-coupled, process-based model components and the authors found strongly declining groundwater levels with declines of up to 10 m close to the Alps in southernmost Germany for their scenario period (2036–2060). Further, several studies investigated future groundwater recharge in different contexts for smaller subregions of Germany using mainly water balance models or process-based models (Barthel

et al., 2012; Herrmann et al., 2016; Kersebaum and Nendel, 2014; Kreins et al., 2015; Neukum and Azzam, 2012; Wegehenkel and Kersebaum, 2009).The application of ANNs to study groundwater level development in the long-term and in the context of climate change for a larger area like Germany has not been performed yet. Related studies with applications of ANNs either used a very small number of wells (Ghazi et al., 2021; Idrizovic et al., 2020; Jeihouni et al., 2019a) and limited time horizons (Ghazi et al., 2021; Jeihouni et al., 2019a) or use ANNs without directly presenting future climate signals to the ANN (Idrizovic et al., 2020). In the case of streamflow runoff simulation, however, ANNs have been successfully applied to analyze the future development under climate change influences in several catchments all over California (Duan et al., 2020) as well as in two catchments in China (Gao et al., 2010; Lee et al., 2020).

In this study, we use a 1D-CNN approach (Wunsch et al., 2021) to build 118 site-specific models, well distributed over Germany in the respective uppermost unconfined aquifer, which are able to predict weekly groundwater levels with high accuracy using only precipitation and temperature as inputs in the past. We visually check the model output plausibility under an artificial extreme climate scenario in the past (Duan et al., 2020) and investigate how the model has learned input-output relationships using an XAI approach (SHAP (Lundberg and Lee, 2017)). We then use the trained CNN models to investigate the future climate-driven groundwater level development for the selected sites, using precipitation and temperature derived from different RCP scenarios (2.6, 4.5, 8.5) (Moss et al., 2008) of bias-corrected and downscaled (5×5 km^2) climate projection data (Brienen et al., 2020) from different climate models. These climate models ("core-ensemble") were preselected by the German Meteorological Service (DWD) to represent 80-90% of the spread of the full ensemble of all available and suitable (according to certain quality criteria) climate projection results under the respective RCP scenario for Germany (DWD, 2018) based on CORDEX-EUR11 (EURO-CORDEX, 2018) and ReKliEs-De (Huebener et al., 2017) (see methods section). As we use purely climatic input variables, we can only project the influence of direct climate change effects, while secondary, most certainly stronger indirect effects, such as increased groundwater pumping, are not included in this study. However, due to high prediction accuracy in the past, the selected sites show a strong relationship between climate variables and groundwater and are unlikely to be under the influence of strong groundwater withdrawals or comparable effects. They are, therefore, suitable for predicting that part of the future groundwater level trend that results from direct climatic influences, as long as the basic input-output relationships remain unchanged.

2 Methods

2.1 Data

We used weekly groundwater level data from 118 different sites, well distributed all over Germany (Figure IV.1a). All wells are located in the unconfined, uppermost (thus mostly shallow) aquifers, which are most likely to be subject to direct climate change effects (Kløve et al., 2014). Greater depths to groundwater are predominantly found in fractured and karstic aquifers. For additional details on the sites, please refer to the supplementary material (Supplementary Table S1). Groundwater level records of all sites show very different lengths (Figure IV.1b), from 15 to 67 years, with a median length of 36 years. Data gaps were closed using the information of several related groundwater level time series with highly correlated dynamics derived from an earlier comprehensive cluster analysis based on hydrograph dynamics (Wunsch and Liesch, 2020; Wunsch et al., 2022b). Alternatively, PCHIP (Piecewise Cubic Hermite Interpolating Polynomial) was used to close smaller data gaps, where no correlated hydrograph information was available. In our dataset, 48 time series had no missing values; another 44 had less than 2% interpolated values. Only very few time series show a higher proportion of interpolated values (11 time series $>$4%). More information on interpolated values can be found online in the released dataset.

Input variables for our models are precipitation and temperature, thus purely climatic. These variables are widely available and easy to measure both in the past and present and are also well evaluated in terms of climate projection output. Precipitation serves as a proxy for groundwater recharge, temperature for evapotranspiration. Additionally, the temperature usually shows a distinct annual cycle, which also provides the models with valuable information on seasonality. Since we specifically selected wells with high forecast accuracy in the past (see Model Calibration and Evaluation), we can assume that the groundwater dynamic at these wells is mainly dominated by climate forcings. As long as no fundamental change of the system relations occurs (e.g., newly installed groundwater pumping or severe changes in land use nearby), we can expect reasonable results for our simulations, as we explore only the influence of changing climate and assume other boundary conditions fixed.

Besides the GWL data itself, we based our analysis on several datasets. The models were trained using temperature and precipitation data from the HYRAS dataset (Frick et al., 2014; Rauthe et al., 2013), which is a gridded (5×5 km^2) meteorological dataset based on observed data from meteorological stations ranging from 1951 to 2015. To evaluate the influence of climate change we used RCP scenario data from several selected climate projections that form the so-called core-ensemble defined by DWD (DWD, 2018) (Figure IV.1). Depending on the scenario and the considered variable, this ensemble represents 80-90% of the ensemble spread

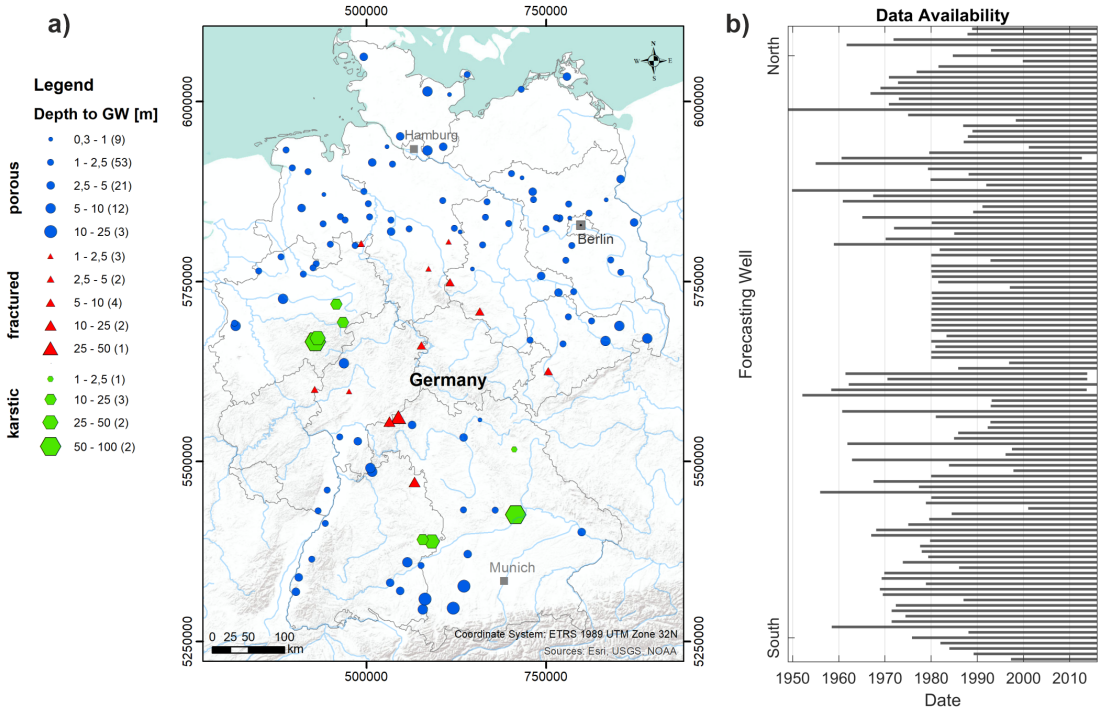


Figure IV.1: **a)** Position, type of aquifer and depth to groundwater for each study site. **b)** Time series length of all study sites ordered in North-South direction.

of the possible climate signal within the larger "reference-ensemble" (DWD, 2018). The latter, in turn, constitutes all available and quality-assessed projections for Germany. Further, we received the projection data bias-adjusted onto the HYRAS dataset and regionalized on a 5×5 km^2 grid by Brienen et al. (2020). For each site, the mean of 3×3 cells around the cell with the respective well was chosen as input for the simulations, following the best practices by DWD to reduce uncertainty resulting from the grid cell size.

Generally, the used climate projections show a slight increase in precipitation sums and a significant temperature increase of several degrees Celsius for Germany by 2100 (EURO-CORDEX, 2018; Huebener et al., 2017; Jacob et al., 2014), more precise values depending strongly on the considered scenario. For RCP8.5, an input data analysis at the relevant 118 sites of this study showed a consistent annual average temperature increase in all regions of Germany of several degrees Celsius (mostly between 3°C and 4°C). Only very slight spatial patterns emerge, with strongest increases in the South (up to 4.7°C) and generally slighter increases in the Northwest, probably due to a buffer effect near the coast. For the total annual precipitation, non-significant changes ($p \geq 0.05$) dominate. The fewer significant changes partly show opposing trends, depending on the projection. One projection shows consistently decreases of mostly -150 mm (max: -367 mm in the far South). Some other projections show increasing precipitation instead (mostly around 100 mm) except for the

Table IV.1: Climate projections used in this study and according abbreviations used throughout the text. For more information on the models please visit www.euro-cordex.net.

	Projections	Abbrev.
RCP8.5	CCCma-CanESM2_rcp85_r1i1p1_CLMcom-CCLM4-8-17	p1
	ICHEC-EC-EARTH_rcp85_r1i1p1_KNMI-RACMO22E	p2
	MIROC-MIROC5_rcp85_r1i1p1_GERICS-REMO2015	p3
	MOHC-HadGEM2-ES_rcp85_r1i1p1_CLMcom-CCLM4-8-17	p4
	MPI-M-MPI-ESM-LR_rcp85_r1i1p1_UHOH-WRF361H	p5
	MPI-M-MPI-ESM-LR_rcp85_r2i1p1_MPI-CSC-REMO2009_v1	p6
RCP4.5	ICHEC-EC-EARTH_rcp45_r1i1p1_KNMI-RACMO22E_v1	p1
	ICHEC-EC-EARTH_rcp45_r12i1p1_KNMI-RACMO22E_v1	p2
	ICHEC-EC-EARTH_rcp45_r12i1p1_SMHI-RCA4_v1	p3
	MOHC-HadGEM2-ES_rcp45_r1i1p1_CLMcom-CCLM4-8-17_v1	p4
	MPI-M-MPI-ESM-LR_rcp45_r1i1p1_MPI-CSC-REMO2009_v1	p5
	MPI-M-MPI-ESM-LR_rcp45_r2i1p1_MPI-CSC-REMO2009_v1	p6
RCP2.6	ICHEC-EC-EARTH_rcp26_r12i1p1_CLMcom-CCLM4-8-17_v1	p1
	ICHEC-EC-EARTH_rcp26_r12i1p1_KNMI-RACMO22E_v1	p2
	MIROC-MIROC5_rcp26_r1i1p1_CLMcom-CCLM4-8-17_v1	p3
	MOHC-HadGEM2-ES_rcp26_r1i1p1_KNMI-RACMO22E_v2	p4
	MPI-M-MPI-ESM-LR_rcp26_r2i1p1_MPI-CSC-REMO2009_v1	p5

Northwest, where almost no increases are visible. The southern part shows the strongest possible increases in precipitation, up to 300 mm. Under RCP4.5, the respective input data reveals no spatial pattern in the case of the temperature. Input data shows spatially consistent increases mostly between 1°C and 2°C. For the precipitation data, non-significant results dominate. However, the few significant changes show a clear spatial pattern and occur mostly in the South and Northwest, ranging mostly around 100 mm; in the eastern part, we see basically no increasing precipitation. Under RCP2.6, non-significant results are dominating. In terms of the temperature data, however, we find a spatial pattern of slight, yet significant increases (0.5°C to 0.8°C) in the North and Northeast, as well as for the Upper Rhine Graben area in the Southwest. Only a few significant results occur for the precipitation, showing decreases of about -100 mm, mostly in the Northwest. The methodology of this input analysis is similar to the trend analysis described in section 2.5. For map and boxplot representations of these analyses, please refer to the Supplementary Figures S3-S8.

2.2 Convolutional Neural Networks

CNNs (LeCun et al., 2015) are commonly used for image recognition and classification (e.g., Cai et al., 2016; Li et al., 2014) tasks but also work well on sequential data, such as groundwater level time series (Wunsch et al., 2021). The CNNs used in this study comprise a 1D-Convolutional layer with fixed kernel size (three) and optimized number of

filters, followed by a Max-Pooling layer and a Monte-Carlo dropout layer, applying a fixed dropout of 50% to prevent the model from overfitting. This dropout rate is quite high and forces the model to perform very robust training. A dense layer with an optimized number of neurons follows, succeeded by a single output neuron. We used the Adam optimizer for a maximum of 100 training epochs with an initial learning rate of 0.001 and applied gradient clipping to prevent exploding gradients. Early stopping with a patience of 15 epochs was applied as another regularization technique to prevent the model from overfitting the training data. Several model hyperparameters were optimized using Bayesian optimization (Nogueira, 2014): training batch-size (16 to 256); input sequence length (1 to 52 weeks); number of filters in the 1D-Conv layer (1 to 256); size of the first dense layer (1 to 256). All models were implemented using Python 3.8 (van Rossum, 1995), the deep-learning framework TensorFlow (Abadi et al., 2015) and its Keras (Chollet, 2015) API. Further, the following libraries were used: Numpy (van der Walt et al., 2011), Pandas (McKinney, 2010; Reback et al., 2020), Scikit-Learn (Pedregosa et al., 2011), BayesOpt (Nogueira, 2014), Matplotlib (Hunter, 2007), Unumpy (Lebigot, 2010–2020) and SHAP (Lundberg and Lee, 2017).

2.3 Model Calibration and Evaluation

We used weekly groundwater level time series data of varying length (Figure IV.1b). To find the best model configuration, we split every time series into four parts: training set, validation set, optimization set and test set. The test set uses always the 4-year period from 2012 to 2016 (Figure IV.2b, s.a. Figure IV.3a for an example, for few sites where the time series ended slightly earlier, we shifted the 4-year test set period accordingly). The first 80% of the remaining time series before 2012 were used for training, the following 20% for early stopping (validation set) and for testing during HP optimization (optimization set), using 10% of the remaining time series each (Figure IV.2b). As target function during HP optimization we chose the sum of Nash-Sutcliffe efficiency and squared Pearson r (compare Wunsch et al. (2021)), the acquisition function is expected improvement. For each model we used a maximum optimization step number of 150 or stopped after 15 steps without improvement once a minimum of 60 steps was reached. Generally, we scaled the data to [-1,1] and used an ensemble of ten pseudo-randomly initialized models to reduce the dependency towards the random number generator seed. For each of the ten ensemble members, we applied Monte-Carlo dropout during simulation to estimate the model uncertainty from 100 realizations each. We derived the 95% confidence interval from these 100 realizations by using 1.96 times the standard deviation of the resulting distribution for each time step. Each uncertainty was propagated while calculating the overall ensemble median value for final evaluation in the test set (2012-2016). We calculated several metrics to judge the

simulation accuracy: Nash-Sutcliffe efficiency, squared Pearson r, absolute and relative root mean squared error, as well as absolute and relative Bias. Note that we calculate NSE with a long term mean GWL before the test set instead of the test set mean value. Please see Wunsch et al. (2021) for more details on calculation as the same approach was used. We use almost exclusively wells, at which the models showed a very high forecast accuracy in the test-set (mostly NSE and R^2 larger than 0.8, compare Figure IV.2a). Some models were included with slightly lower accuracy (at least NSE and R^2 larger than 0.7) to improve the spatial coverage resulting in a set of 118 wells from all over Germany. For additional details on the error measures and hyperparameters for all sites please refer to our supplementary material. Figure IV.3a shows the model evaluation on the test set exemplarily for one well.

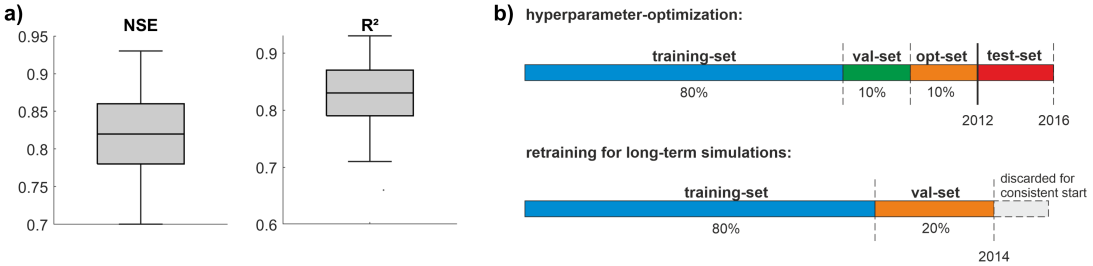


Figure IV.2: **a)** Model performance of all models for the test-set (2012-2016). **b)** Time series splitting scheme for optimization (upper) and retraining (lower).

2.4 Model Plausibility and Interpretability

To perform groundwater level simulations until 2100 we retrained all models using the defined hyperparameters and all data until 2014. Hence, we split the time series only in two parts: 80% for training and 20% for early stopping (Figure IV.2b). Afterwards, we assessed the model stability and the plausibility of the output values in the extrapolated regime accordingly to Duan et al. (2020) by evaluating the model output using artificially altered input data based on historical observed climatology with quadruple precipitation and systematically 5°C higher temperature (Figure IV.3b). As long as the model output does not "blow up" or produce meaningless outputs (Duan et al., 2020), we can hereby improve confidence in the simulation results when investigating the different RCP scenarios. Models showing implausible behavior in preliminary analyses were not considered for this study. We additionally applied a XAI approach to check whether the models have learned correctly in terms of our conceptual understanding of hydrogeological processes. We calculated SHAP (Lundberg and Lee, 2017) values that explain the influence (sign and strength) of every input feature value on the model output (Figure IV.3c). Generally, our models showed that the relationship between input and output was captured plausibly. For example, high precipitation inputs (P, red) produce high

SHAP values and therefore have a strong positive influence on the model output, which corresponds to our basic understanding of the influence of recharge, leading to increasing groundwater levels. Low or no precipitation (P, blue) has a comparably slight negative influence on GWL, whereas high temperature inputs (T, red) have a strong negative influence on the model output. Again, this corresponds with our basic understanding of the governing processes, where high temperature usually means high evapotranspiration, which causes less recharge or even direct groundwater evaporation in some cases. This sounds trivial, however, during preliminary work for this study, we found that not all models capture these relations correctly, which also partly caused erroneous values in the extrapolated regime (see above). Such models were excluded for the final study. Figure IV.3 exemplarily summarizes the model evaluation (a) and plausibility checks (b,c) for one well. Respective figures of all other sites are provided in the supplement (Supplementary Figures S9-S126).

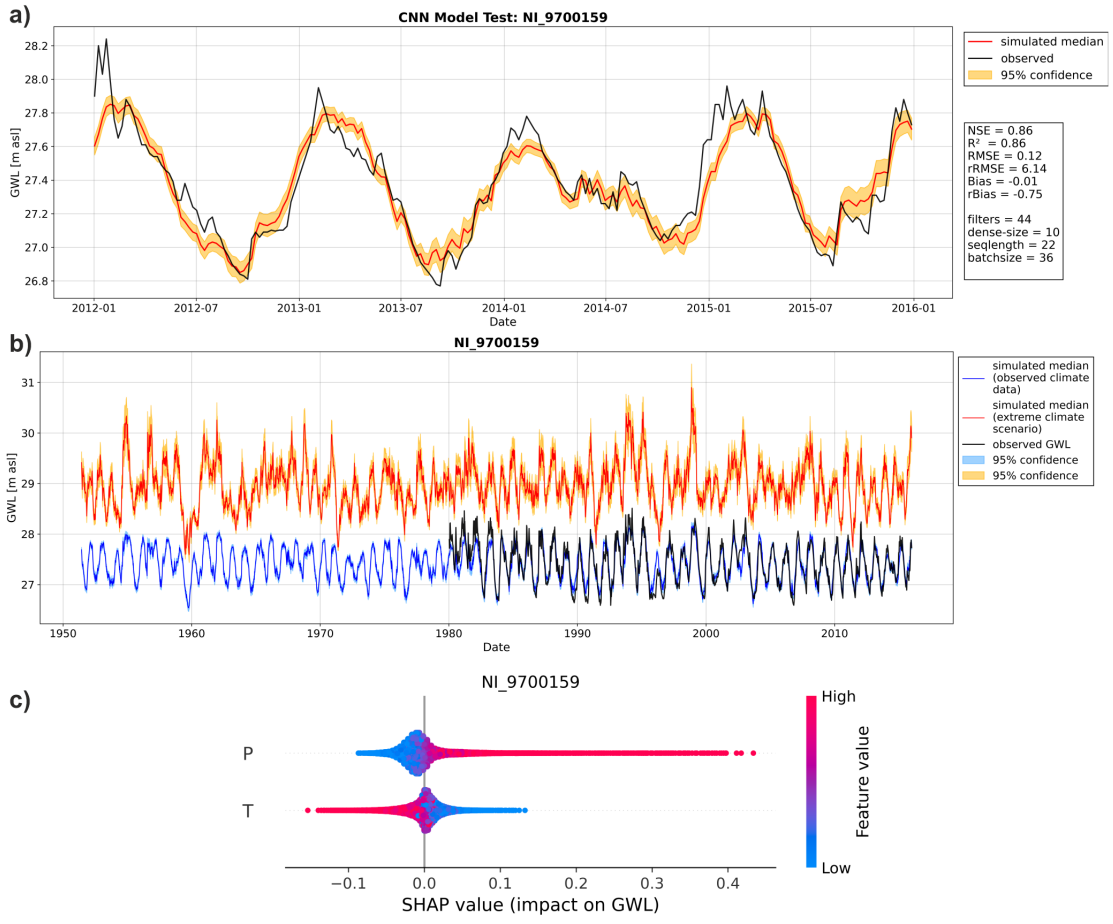


Figure IV.3: **a)** Optimized model evaluation in the past for the test set (2012-2016). **b)** Model output under an artificial extreme climate scenario in the past. **c)** SHAP Summary plot.

2.5 Evaluation of the Projected Groundwater Levels

For our simulation results until 2100, we examined the relative development of the mean as well as the 2.5% (lower extreme), and 97.5% (upper extreme) quantile. All were site-specifically calculated on a yearly basis for each individual projection followed by a linear trend analysis based on Mann-Kendall and Theil-Sen slope. In doing so, we are able to capture both the range and the individual development of all considered future climate projections. Even though considering yearly values, we applied 3PW prewhitening method (Collaud Coen et al., 2020) (implemented in the *mannkendall/Python* (Vogt, 2021) package) to eliminate remaining first-order autocorrelation before applying Mann-Kendall test and calculating corresponding Theil-Sen slopes. To make comparisons between different sites possible, results are normalized on the individual range of each historic groundwater level time series between the years 2000 and 2014 (start of simulation due to data availability). Even though all climate projections are bias-adjusted on the HYRAS training dataset, they still do not depict the real climate development for individual years (also historically), which can cause a bias between the end of historical data records and the start of our simulations. We, therefore, investigated the trend of the aforementioned quantities between the start of the simulation and the end in 2100 and did not directly consider the end of the historical records. We examined each groundwater level development using Mann-Kendall linear trend test (Hussain and Mahmud, 2019) and derived the relative development in percent from a linear fit using Theil-Sen slope (Sen, 1968). For Mann-Kendall test, we considered a trend significant for $p<0.05$, and we further provide upper and lower 95% confidence bounds of the Theil-Sen slopes for all significant trends.

3 Results

3.1 Individual Projection Results

For each of the examined 118 test sites, we simulated the future weekly groundwater level development based on five to six climate projections per RCP scenario. Since these climate projections differ considerably in detail for individual future time periods, we also obtained several different future groundwater level simulations per scenario and considered site, which should only be interpreted based on longer time periods (at least 30 years) (Kreienkamp et al., 2012), such as with a linear trend analysis performed here considering the whole time period of more than 80 years. Figure IV.4 depicts the results of our analysis for RCP8.5, in terms of the relative change in % between the start (2014) and the end of the simulation period (2100) for each of the six projections under RCP8.5 for: (a) the annual mean, (b) the

annual upper extreme (97.5%) quantile and (c) the annual lower extreme (2.5%) quantile. For each site, all displayed developments are ordered by the strength of the change, which does not necessarily correspond to the numbering of the projections (Figure IV.1). The given boxplots in Figure IV.4d provide more detailed information on the three maps, as well as confidence intervals on the statistical analysis. The values of the non-significant trends are not shown in the boxplots, which has to be kept in mind for interpretation. For detailed numbers on the boxplots, we refer to Supplementary Table S3.

In the case of the annual mean, approximately 47% of all simulations (332 of 708, i.e., six projections for each site) show a significant trend until 2100. There is always at least one result for each site significant ($p<0.05$), which, however, also means that there are several sites with mainly non-significant trends (gray). The large majority of the significant trends is negative, with a median ranging between -18% (p1) and -6% (p6). Note that the uncertainty (shown by the boxplots in Figure IV.4d) can be quite high from the trend analysis alone, and we further see that the lower bound sometimes shows a larger spread, thus a higher uncertainty, than the upper bound. In Figure IV.4d, we also observe that p1 systematically shows the strongest declines until 2100, being significant for 114 of the 118 wells. The overall maximum decline of the annual mean is -35%, clearly indicating the different character of p1 compared to the other projections. Especially p3-p5 show more moderate changes of the mean (median ranges from -8% to -11%), with many non-significant trends (50%-58%). Simulations based on p2 and p6 only find significant trends for 22% and 29% of all sites, respectively, and additionally are moderate in their significant results. Three projections (p2, p3, but mainly p6) even show some positive trends until 2100, however, overall, these are rare and occur at sites, where other projections simultaneously show at least non-significant or even negative trends. In absolute numbers, the median changes are in the order of -0.1 m to -0.3 m, which is highly dependent on the individual groundwater level range at each site. Despite many non-significant and some positive trends, there is a clear tendency of declining mean groundwater levels until 2100. Additionally, we can observe a slight spatial tendency with more and stronger significant negative trends in some areas of northern and eastern Germany, where we also find the strongest overall relative declines. In southern Germany, many wells show multiple non-significant trends and most of the positive changes are also scattered in this region; however, some of the southernmost wells also show some very strong decreases for single simulations.

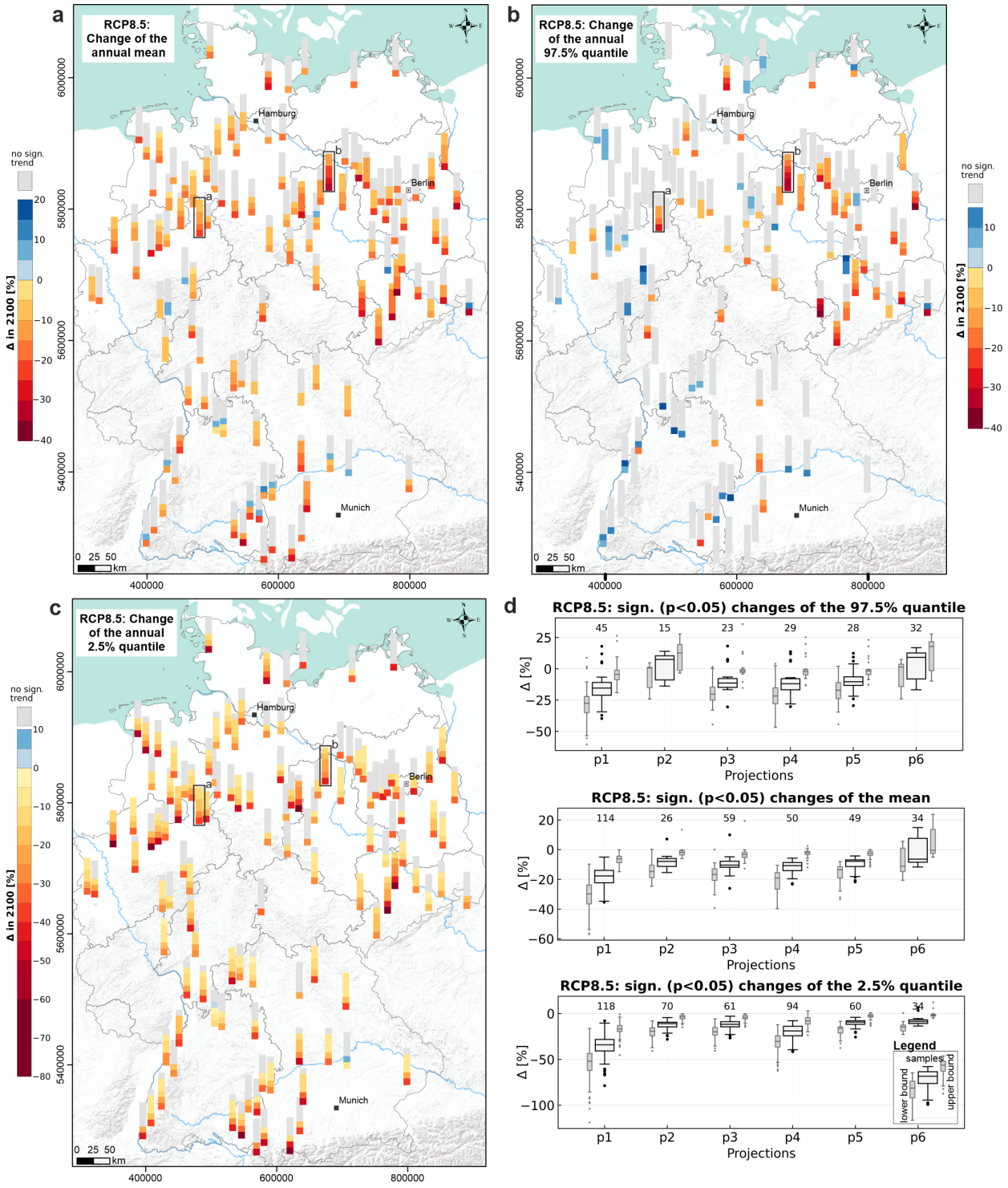


Figure IV.4: Change of groundwater levels [%] in 2100 relative to 2014 (start of sim.) for each site and each climate projection, based on a linear trend analysis: **a)** mean, **b)** 97.5% quantile, **c)** 2.5% quantile; the order corresponds to the strength and sign of the change. **d)** Boxplots showing the significant changes for a-c, light gray/sideways boxplots show the uncertainty of the change as 95% confidence interval. Numbers above boxplots depict the sample size (significant trends). Black boxes on maps depict the sites shown in Figure IV.6

The results for the upper extreme value quantile (97.5%) confirm these spatial patterns partly. In Figure IV.4b we clearly observe many significant declines in eastern Germany, while the large majority (76%) of the trends in whole Germany is considered to be non-significant. Increasing trends are found comparably often, with increases close to 18% (p1, p3, p6). Comparing the projections (Figure IV.4d), we find a similar behavior as before: p1 shows the strongest significant decreases (down to -40%, conf.-interval: -61% to -19%), p3, p4, and p5 tend to move in the moderate negative range (medians around -11%), while p2 and p6 more often show positive trends (positive medians of the significant trends). Particularly the latter cause a partly contradictory development of the upper extreme values compared to the mean. The absolute numbers of changes are again in the order of few tens of centimeters.

The tendency of declining groundwater levels we observed for the mean gets clearer for the lower extreme values (2.5% quantile) shown in Figure IV.4c. We still observe 38% non-significant trends, however, the remaining 62% show almost exclusively negative changes with a maximum decline of -79%. The median change of the 2.5% quantile of all projections ranges between -34% for p1, which again shows the strongest declines, followed by p4 (-19%), as well as p2, p3, p5, and p6 with a median change around -9% to -12% each (lower bound: -20%, upper bound: -2%). The latter four, and especially of them p6, contain the majority of non-significant trends, the changes shown in the boxplots, therefore, tend to be overestimated. There are only few sites where only one result is considered significant. These occur, for example, near the Baltic Sea coast as well as the central and eastern part of northern Germany. Quite strong relative decreases are visible in eastern Germany and in the western part of northern Germany as well as at the southernmost sites. This pattern is largely consistent with the spatial pattern of the mean mentioned above. When translating into absolute units, most median decreases (p2-p6) are in the order of -0.1 m to -0.4 m. For p1 and when considering the annual lower extreme value quantile, the median decrease reaches even -0.6 m. From all projections except p6, we see that of all significant changes for the 2.5% quantile, at least a decrease of -0.1 m is observed (summarized in Supplementary Table S3).

The spatial patterns in Figure IV.4 (a-c) are particularly interesting because they do not intuitively follow from the patterns of the input data (compare Figures S7 and S8). Considering all results of RCP8.5, we see a clear tendency toward declining groundwater levels overall, with stronger declines for lower quantiles, i.e., groundwater level lows will occur more frequently and will be more severe in the future. At the same time, except for East Germany, mostly no or even increasing trends are found for upper extreme values, which means that the overall variability will increase considerably by the end of the century.

Figure IV.5 summarizes the results for the other considered RCP scenarios 2.6 and 4.5. For the former, which is a stringent mitigation scenario in terms of greenhouse gas emissions, we see that generally the number of significant samples ($p<0.05$) in total is low, with only 6% to 8%, depending on the quantile considered. We generally see smaller decreases compared to RCP8.5; the upper extreme value quantile does no longer show considerable positive changes. Supplementary Figure S1 shows the spatial distribution of the found changes. We can detect no spatial pattern for the 2.5% quantile, but (slight) decreases all over Germany, dominated by mostly non-significant results. The mean and the 97.5% quantile, however, show that decreasing changes occur preferably in northern Germany, whereas the southern part either shows few slight decreases for the mean or remains mostly non-significant for the upper extreme values. The results strongly indicate that the reduced greenhouse gas emissions of the RCP2.6 scenario also translate to a distinctly reduced impact on the groundwater level development, especially compared to the opposite RCP8.5 scenario. Nevertheless, decreasing trends are still visible all over Germany, showing that even for RCP2.6 with a limited global warming below 2°C compared to pre-industrial temperatures, a change in water availability is to be expected.

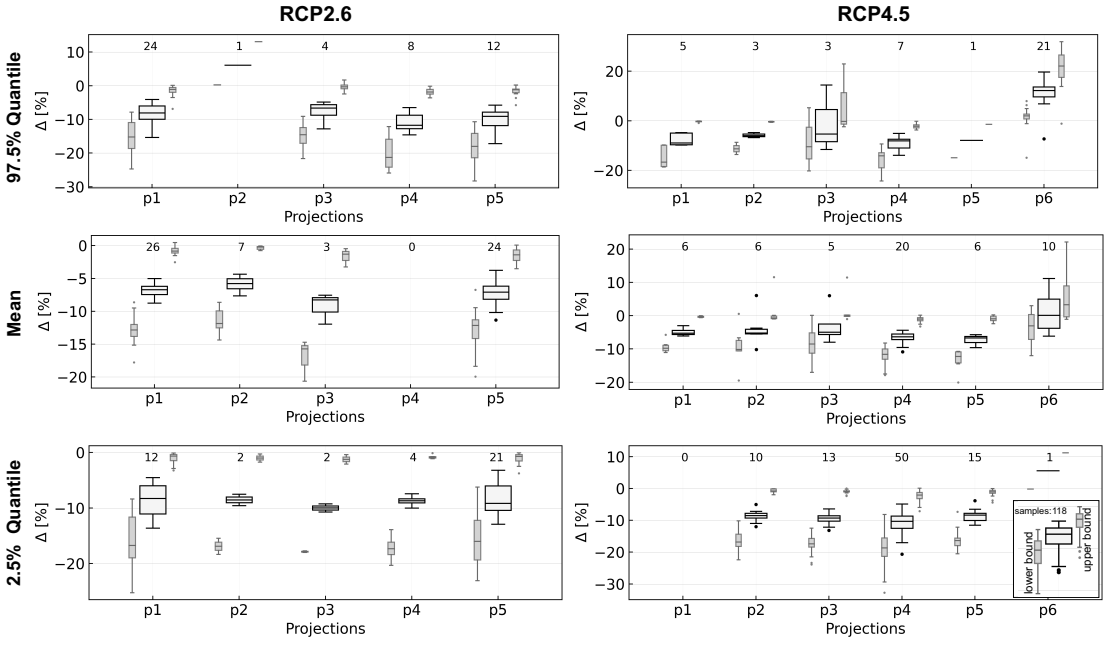


Figure IV.5: Boxplots showing the significant ($p<0.05$) relative changes between 2014 and 2100 based on linear trend analyses of annual quantiles (2.5% and 97.5%) and the annual mean under RCP2.6 (left) and RCP4.5 (right). Light grey/sideways boxplots show the uncertainty of the change as 95% confidence interval. Numbers above boxplots depict the sample size (number of significant trends).

For RCP4.5, changes are also only rarely significant (Q97.5: 6%, mean: 7%, Q2.5: 13% of all samples). Projection p6 represents definitely an increasing groundwater scenario for the future, whereas p1 to p5 mostly show decreases for the significant changes. Except p6, we, therefore, see median changes of all three annual quantiles between -5% and -10%. RCP4.5 and RCP2.6 do not differ here very strongly, but the number of significant samples is a bit higher for RCP4.5 as well as the confidence intervals shown in Figure IV.5 are slightly narrower than in RCP2.6. Differences get clearer spatially, where we find more distinct patterns in the case of RCP4.5 (Supplementary Figure S2) with increasing values almost exclusively in southern Germany (97.5% quantile, less frequent also for the mean). This clearly coincides with the spatial pattern of increasing precipitation in the input data (Supplementary Figures S5-S6). While decreasing changes can be found in northern Germany for the 2.5% quantile, this is less pronounced for the annual mean and even lesser for the 97.5% quantile. For both, sites with exclusively non-significant changes increasingly dominate. For both RCP2.6 and 4.5, we do not find the strong decreasing trends in eastern Germany seen for RCP8.5, however, both scenarios indicate that a stronger tendency of decreasing trends in the North, a slight increasing tendency of upper extreme values for the South, as well as an increasing overall variability (decreasing lower quantiles, constant or increasing upper quantiles) are possible. While for RCP2.6 we do not see that the lower extreme values decrease stronger than other parts of the hydrographs as under RCP8.5, this pattern emerges under RCP4.5 in agreement. Overall, due to the high number of non-significant results, RCP2.6 and RCP4.5 results should be interpreted carefully. Maps, as well as detailed numbers on the boxplots in Figure IV.5, are part of the electronic supplement (Figure S1-S2, Table S4-S5).

Figure IV.6 shows exemplarily the detailed development at two arbitrarily selected sites (black boxes in Figure IV.4) under RCPs 4.5 and 8.5, which, as explained, are the most relevant given the current situation. The simulation results are depicted as time series plots for the far future (2071-2100) and as heatmaps with years as rows and weeks as columns for each of the projections. Heatmaps of both scenarios share the same color scale per site. Heatmaps and time series plots of the simulation results of all other sites and for all RCPs are available online (Wunsch, 2021). The time series plots show the diverging development of some projections in the far future, however, there is no strict sequence of projections in terms of absolute groundwater height, the order can change throughout the years. Most heatmaps visualize the development described above by displaying generally declining groundwater levels (more and darker red, as well as lighter or constant blue shadings towards 2100 in the lower part of the heatmaps). Moreover, we observe increasing lengths of periods with low groundwater levels (wider red shadings) throughout the year. In accordance, wet periods usually get shorter (narrower blue shadings) or even change to red (e.g., in b, RCP8.5, p1, p3, p4). The absolute height of groundwater levels during wet periods does not necessarily decrease but

can even show the opposite behavior (darker blue, e.g., in a, RCP8.5, p6). Most importantly, in both scenarios and at both sites, we can also recognize successions of several dry years. Such periods are visible in the time series plots, but more clearly as dark red horizontal stripes in the heat maps. These are especially critical because drought effects accumulate and dependent ecosystems cannot recover but are instead particularly vulnerable to further damage in subsequent years due to reduced resilience. Although the results should not be interpreted over shorter periods of time (i.e., they do not reflect the absolute timing of an event), they definitely show the increasing probability of such longer-term droughts in the future, especially in the second half of the century.

3.2 Average Projection Results Under RCP8.5

In Figure IV.7 we consolidated the separate projection results under RCP8.5 for each site into one by calculating the mean of the significant trends shown in Figure IV.4. Only sites with at least four (thus the majority) significant projection results are included; the rest is depicted as not significant on average. This is one reason for neglecting RCPs 2.6 and 4.5 in this analysis step, as barely sites with four or more significant results were found there. Another reason is that, at least for the near future, the results of RCP8.5 can be considered most relevant, as it is the scenario closest to our current situation (Schwalm et al., 2020). Even though we investigate a longer time period until 2100, tendencies should be nevertheless useful to estimate near-future developments. The development of the mean is depicted in the upper left map (a), and according to the aforementioned definition, about 30% of the wells (35 of 118) are considered significant on average and on median show a change of -12%. We do not find any wells with increasing mean trends on average and observe a similar spatial pattern as before with the strongest decreases in eastern Germany. For wells in southwestern Germany, we observe a noticeable number of non-significant changes. Overall, we simulated significant absolute average decreases between -0.2 m to -2.1 m for about 18 wells and at least a decrease of -9 cm for all 35 wells in Figure IV.7a. In the case of the annual 97.5% quantile, the consolidated results show mainly no trends, especially for southern Germany. Two sites in northern Germany are expected to show increased upper extreme values up to a maximum of 7.5% or 0.2 m, however, we still observe a spatial pattern of decreasing upper extreme values in eastern Germany up to -24%. Hence, in this area, the groundwater levels possibly decrease in every part of the annual cycle and with comparably high certainty (many consistent significant simulations). This also applies to the lower extreme values (2.5% quantile) that show on average significant decreases for more than half of the examined sites, with median decreases of -17% (or -0.3 m) (compare Figure IV.7d, e). On this map, no clear spatial pattern is recognizable any longer.

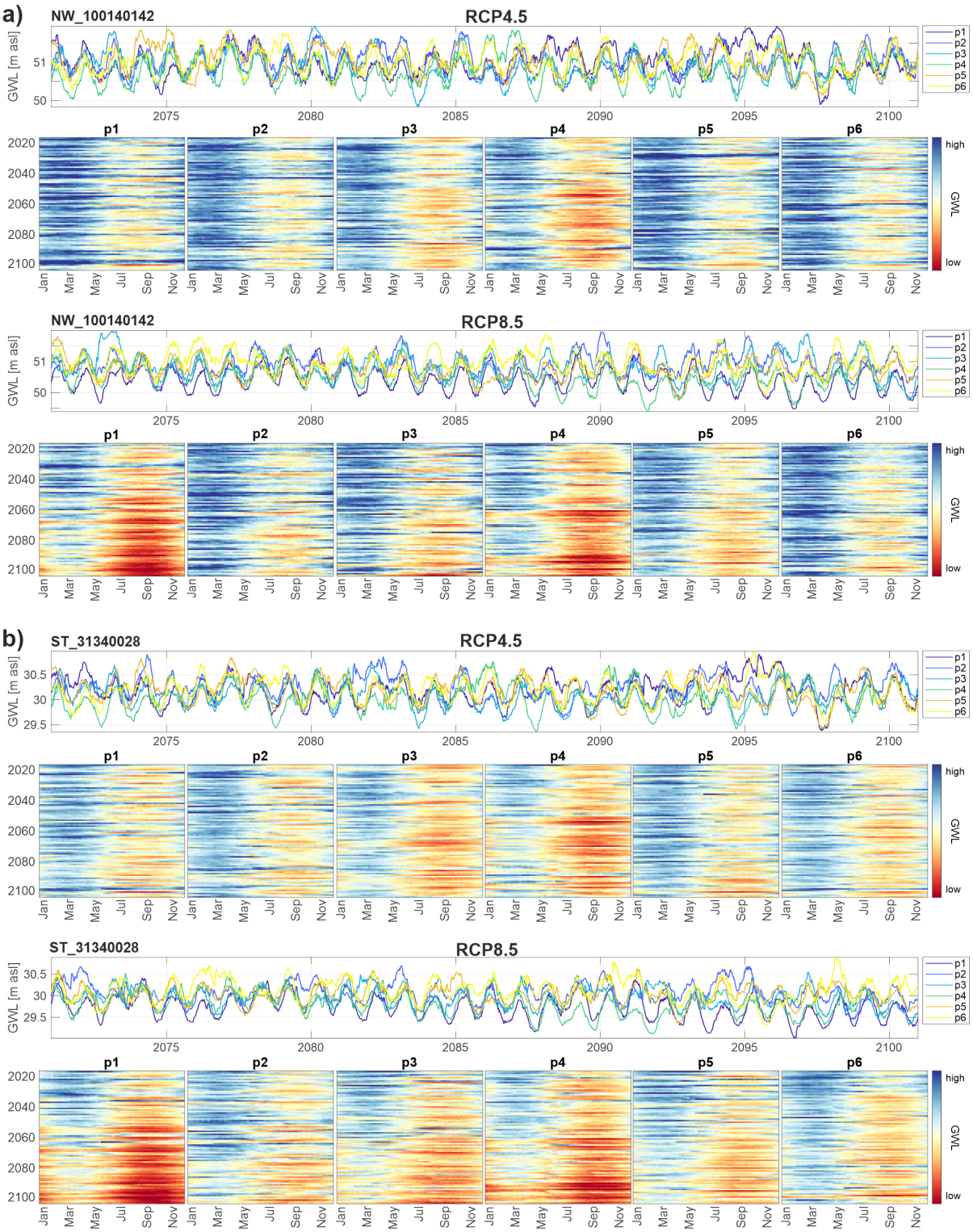
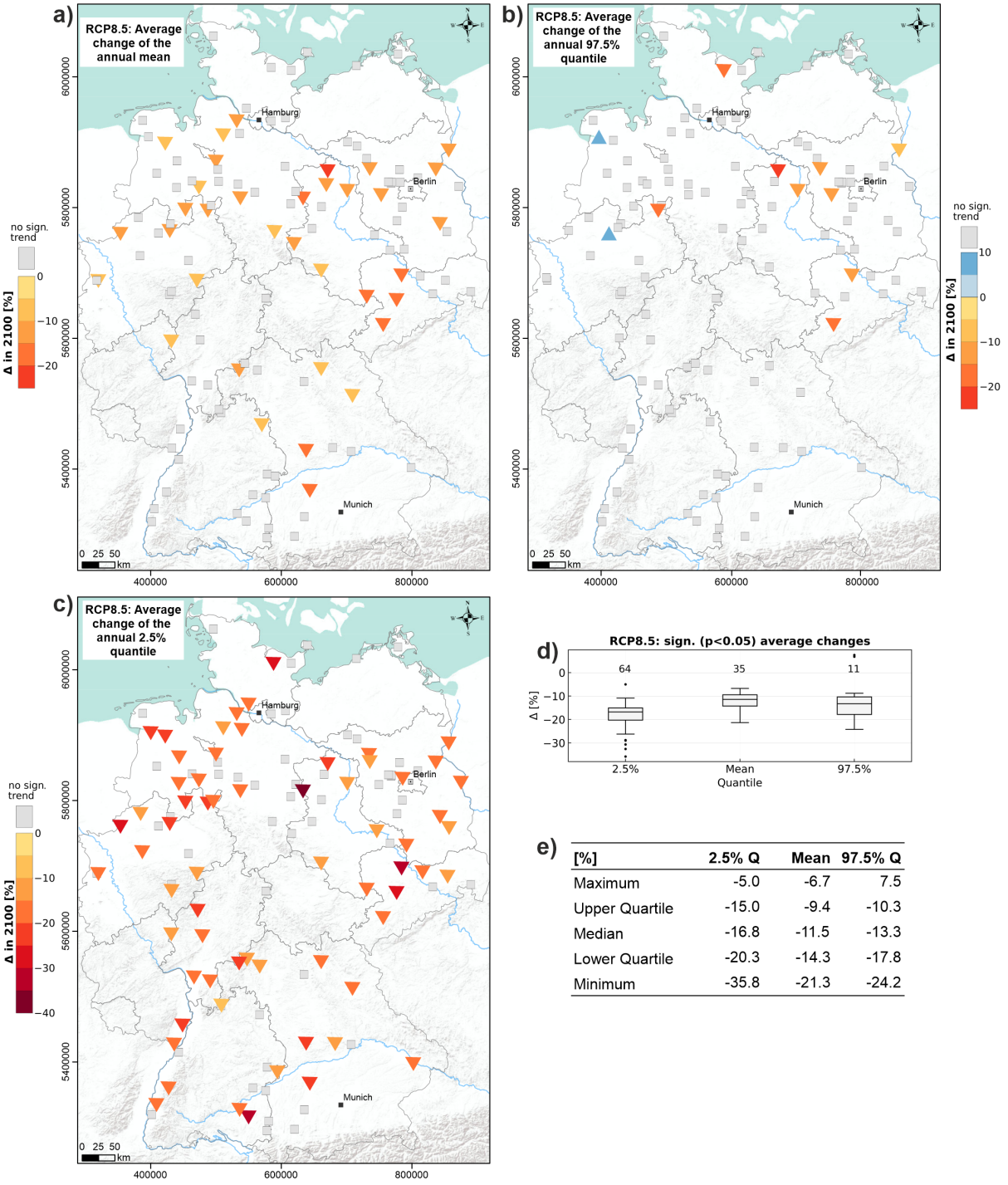


Figure IV.6: RCP4.5 and RCP8.5 results for two arbitrarily selected sites marked by black boxes in Figure IV.4 **a)** NW_100140142, **b)** ST_31340028. Heatmap plots show the whole simulation period for each of the projections under each of the considered scenarios. Columns of each plot as weeks during the year and rows as the year (top: 2014 – bottom: 2100).



e)

[%]	2.5% Q	Mean	97.5% Q
Maximum	-5.0	-6.7	7.5
Upper Quartile	-15.0	-9.4	-10.3
Median	-16.8	-11.5	-13.3
Lower Quartile	-20.3	-14.3	-17.8
Minimum	-35.8	-21.3	-24.2

Figure IV.7: Averages for all sites of the significant trends (at least four) of the **a)** annual mean, **b)** the annual 97.5% and **c)** the annual 2.5% quantiles shown also in Figure IV.4. **d)** Associated boxplots and **e)** the corresponding table.

3.3 Model Input Analysis

From the combined analysis of our groundwater level simulations, especially under RCP8.5, and the model inputs presented in the data section and Supplementary Figures S3-S8, we conclude that for shallow aquifers temperature is the main driving factor for declining groundwater levels, rather than precipitation. This applies because mostly non-significantly changing or even increased precipitation is projected, however, our models still frequently show declining groundwater level tendencies. Therefore, these are most likely caused by the significantly increased temperature until the end of the century. Nevertheless, especially under RCP4.5, spatial precipitation data patterns from the input data translate into related patterns of groundwater levels, which shows the also high importance of precipitation. Our results are consistent with other studies, which indicate that the reduction in water availability in the future is driven primarily by changes in temperature (Thober et al., 2018). This is also reflected in the results of the model interpretability approach (SHAP values (Lundberg and Lee, 2017)) that we used to check the plausibility of our model outputs. The minimum SHAP value for T is mostly lower than the minimum SHAP value observed for P (except for eight sites); i.e., the models have learned that high temperatures can cause stronger decreasing groundwater levels than low precipitation. This is, however, only an interpretation of what was learned, which agrees with our conception. Causality cannot be derived from this.

3.4 Sources of Uncertainty

There are different sources of uncertainty in our study. Considering the groundwater model itself, there exists uncertainty directly from different model realizations as well as the uncertainty due to limited model precision. The former was derived from a Monte-Carlo dropout approach and is on average consistently small for all models (orange sections Figure IV.3a and Supplementary Figures S9-S126), the latter is hard to generalize, as it is different for each site. However, we only used models with high performance in the past, checked the conceptual correctness of what was learned using SHAP values, and investigated the stability of the model output in the extrapolating regime, to improve the confidence in the model simulations. However, it is important to mention that data-driven models generally have difficulties in predicting extreme values. Figure IV.8 shows the yearly relative model bias on different quantiles during the model testing period (2012-2015, normalized on the historic min-max range of each individual time series). On average, the models show a very small bias; however, a considerable bias occurs for extreme values (2.5% and 97.5% quantiles). Lower extremes are overestimated by 4.8%; upper extremes are underestimated by 9.6% (both on median). Thus, the analyses of future extreme values are less robust than for

the mean. Nevertheless, we argue that (i) reasonable conclusions can still be derived from relative trends and tendencies even for the extreme values at each site and (ii) since the extreme values are underestimated, the analyses constitute a kind of best-case scenario.

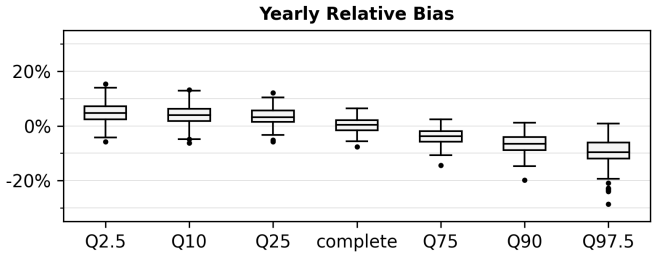


Figure IV.8: Evaluation of the yearly, relative model bias on different yearly quantiles at all sites for the four-year model testing period.

Concerning the simulation of climate change impact, we are not extrapolating in a classical sense. Mean values and frequencies of input values change in the future, but the total range of these values is usually already present in the training data. Scaling uncertainty due to the differences between a single location and the grid cell sizes are certainly present, however, by achieving high performance in the past using training data in the same grid resolution, we can assume that this influence is not severe. To account for atmospheric process scales in the climate models that are not reliably scaling down to cell resolution, we follow the DWD best practice recommendation of considering 3x3 cells rather than one cell that best matches the site location. Regarding the uncertainty deriving from climate models or the considered scenario themselves, we consider different RCP scenarios, each based on a whole ensemble of individual climate models. Finally, the uncertainty from the applied statistical tests (Mann-Kendall test and Theil-Sen slopes) is directly communicated in the text and figures.

4 Discussion

The results of our simulations show a nationwide decrease in climate-driven groundwater levels by the end of the century under the RCP8.5 scenario. The results for RCP2.6 and RCP4.5 show comparably few significant changes, thus have to be interpreted with care in absolute and relative numbers. However, this also means that mitigation of greenhouse gas emissions could have a visible effect, at least for the climate-driven part of the total future groundwater levels in Germany. Nevertheless, even for RCP2.6, decreases in all considered quantiles were found all over Germany for some projections. We, therefore, will probably have to cope with drought effects and changing water availability in any of the investigated scenarios, especially because current estimations of future climate change impacts (UNFCCC,

2021) still exceed the RCP4.5 scenario. Especially for the near future, the results under RCP8.5 are most relevant (Schwalm et al., 2020) because its path is closest to our current situation.

The absolute changes even under RCP8.5 may seem small, but the facts that we investigated almost exclusively shallow aquifers and sites with comparably small depths to groundwater reinforce the importance of the results, predominantly in terms of water availability for vegetation and agriculture. A decline of several tens of centimeters (depending on the projection and the area) can be vital for plants during hot and dry periods if, as a result, the groundwater is no longer accessible. Furthermore, a related study showed that for large parts of northern Germany, a decline of the groundwater levels by 10 cm can be considered critical in terms of altered streamflow discharge due to reduced baseflow from groundwater (de Graaf et al., 2019). This has already been visible during the summers of 2018-2020, when simultaneously to low groundwater levels, also extremely low water levels in surface waters were observed (Wriedt, 2020). Our results show a clearer tendency of declining groundwater levels in the North and the East compared to the South (Figure IV.7a), which emphasizes the already existing trends and patterns. However, in the southernmost part of Germany, for some individual projections, we also find some of the strongest declines (Figure IV.4). It is very important to note that the assessed results are only long-term averages of a future development. As recent developments illustrate, the succession of several dry years is much more critical than the overall trend. In such periods, the projected effects accumulate over consecutive years to extremely low groundwater levels, and thus more severe consequences are to be expected. Such longer dry periods are most likely to be averaged out in a linear trend analysis, as performed in this study. Nevertheless, we see an increasing frequency of them in all RCP scenarios (Wunsch, 2021), especially in RCP8.5 and less pronounced in RCP4.5 (Figure IV.6). Future research should pay attention to this aspect more intensively. It is also important to highlight that we only model direct climate effects on groundwater levels, and we assume that the basic input-output relationship or system behavior does not change. However, it can most certainly be expected that the system behavior will be influenced by future changes in groundwater extractions, changes in vegetation and land use, as well as surface sealing and other related factors. Groundwater withdrawals, in particular, are expected to increase due to (i) the regionally growing population, especially in metropolitan areas (drinking water demand) and (ii) the increasing demand for industry, energy, and especially irrigated agriculture. As a result, the groundwater level will inevitably drop further if no active measures such as limitation of withdrawals, avoidance of irrigated agriculture by changing crop types, or even artificial recharge by infiltration, are applied. Despite all these limitations, the results give a good impression of the magnitude of changes to be expected purely due to direct climatic influences.

Acknowledgments

Open Access funding enabled and organized by Project DEAL. We acknowledge the support and advice by the German Meteorological Service in providing and handling the climate data.

Chapter V

Karst Spring Modeling

This chapter is based on a study published in Hydrology and Earth System Sciences (HESS) and is an edited reprint of:

Wunsch, A., Liesch, T., Cinkus, G., Ravbar, N., Chen, Z., Mazzilli, N., Jourde, H., Goldscheider, N., 2022. Karst spring discharge modeling based on deep learning using spatially distributed input data. Hydrology and Earth System Sciences 26, 2405–2430. doi:10.5194/hess-26-2405-2022

The following links provide access to the associated online resources. This study does not contain any electronic supplementary material but an appendix, which is presented directly at the end of the chapter.

Paper

Code

1 Introduction

Karst aquifers and karst springs are crucial for freshwater supply in many regions, and 9% of the global population partly or fully rely on karst water resources (Stevanović, 2019). Karst systems, in general, are characterized by high structural heterogeneity due to the at least in large parts unknown conduit network, which controls the highly variable groundwater flow. These factors make modeling difficult. Nevertheless, different approaches exist, which Jeannin et al. (2021) classify as hydrological models (fully distributed models), pipe flow models (semi-distributed models), and data-driven models (including reservoir models). ANNs or its subgroup of DL models are part of the last group. In contrast to the other two categories, which usually require detailed system knowledge in order to achieve high-quality results, DL approaches offer an alternative possibility of modeling by being able to establish an input-output relationship automatically, without detailed system knowledge necessary. Even though ANNs are not a standard method in karst modeling yet, different types of ANNs have been applied in modeling karst water resources for quite a long time. As one of the first applications Johannet et al. (1994) showed that karst spring discharge modeling is possible with ANNs. Since then, application of ANNs in hydrology in general received ever-growing research attention (e.g., Maier and Dandy, 2000; Maier et al., 2010). This has amplified even more in the last years, mainly because of the recent success of DL models (e.g., Kratzert et al., 2018). Rajaee et al. (2019) more recently reviewed applications of ANNs on groundwater; Sit et al. (2020) summarize applications on hydrology and water resources in general. Recurrent neural networks, such as LSTM (Hochreiter and Schmidhuber, 1997) are standard models for time series modeling because they possess explicit or implicit memory to remember past time steps, which helps to infer the future. A consequence is that they are trained sequentially, which can make them computationally expensive. CNNs (LeCun et al., 2015) on the other hand, use convolutions along the time axis (1D-CNNs) to learn temporal features and can be trained batch-wise, which therefore usually makes them computationally favorable over RNNs. One example for this fact exists in the related domain of groundwater level forecasting, where Wunsch et al. (2021) showed that 1D-CNNs are considerably faster than RNNs in the case of single-site model application. CNNs, at the same time, exhibited stable results through a comparably low dependency on the random network initialization and achieved some of the highest performances in this specific study (better than LSTM). Other authors similarly applied CNNs successfully for either GWL forecasting (Afzaal et al., 2020; Lähivaara et al., 2019; Müller et al., 2020) or rainfall-runoff modeling (Hussain et al., 2020; Van et al., 2020). Müller et al. (2020) find in contrast to Wunsch et al. (2021) that CNNs take a considerably longer time to optimize than LSTMs, yet both studies agree that they outperform LSTMs in terms of accuracy. Given these favorable properties of CNNs, we choose 1D-CNNs for karst spring discharge modeling for our study. To our best knowledge

Jeannin et al. (2021) is the only study yet, applying CNNs for karst spring discharge modeling in some first experiments, and they also find CNNs to be superior over LSTMs in terms of testing performance.

Data-driven approaches, in general, are considered to be black boxes. A way to still build confidence in a model's decisions is to understand what the model is doing (ideally, even why) by using XAI approaches. There are different techniques that are potentially suited for this purpose, depending on the specific goal. Such approaches are not only useful to gain trust but also help during model building to debug the model and to check what aspects it is focusing on (McGovern et al., 2019). The class of wrapper methods (Kohavi and John, 1997) incorporates both the data and the trained model to interpret what a model has learned. Methods from this class are, for example, impurity importance for determining feature importance in random forest (RF) models (Louppe et al., 2013), permutation importance (Breiman, 2001) both for RF and DL models, and partial dependence plots (Friedman, 2001) that also reveal why a predictor is important. See McGovern et al. (2019) for an overview on these and several other model interpretation and visualization methods. Especially for image-alike data, input sensitivity approaches seem suitable, as focus regions of the model on the image can be visualized. Two well-known approaches are occlusion sensitivity (Zeiler and Fergus, 2014) and RISE (Randomized Input Sampling for Explanation) (Petsiuk et al., 2018). Both methods show how relevant each pixel or area is for the decision of the model (image classification) and can generate an importance heatmap (saliency map) for visualization. The idea behind both algorithms is to use masked versions of an input image and by obtaining the respective model output to learn the focus regions. A very closely related approach to generate a saliency map was recently proposed by Anderson and Radić (2022), which in contrast to RISE and occlusion takes the physical meaning of the absolute value of each variable at each pixel into account during the perturbation of the input data.

One drawback of the 1D-CNN approach, as well as most other data-driven approaches, is the dependency on high data availability and quality. However, climate stations are often not available within the catchment itself, do not match the data availability of the discharge time series (period or temporal resolution), or are more distant and thus do not truly represent the climatic conditions within the catchment. Gridded climate data can provide a solution to such data availability problems. Several openly available products exist (e.g., ERA5-Land (Muñoz Sabater, 2019), E-OBS (Cornes et al., 2018)), which provide climate data for several decades and with, in terms of karst spring modeling, appropriate temporal (hourly or daily) and spatial ($0.1° \times 0.1°$) resolution. However, especially for karst springs, it is not straightforward to extract relevant time series from the gridded data, because the spatial extent of the grid cell containing the location of the spring usually does not coincide well with the associated spring catchment position. Moreover, especially for karst springs, the

catchment is often not well-known and, for larger springs, can stretch over several grid cells. If the exact position of the catchment is unknown, using gridded data has the advantage that a broader region can be taken into account as input to let the model learn the relevant grid cells automatically.

Besides such modeling aspects, the delineation of karst catchments is generally important to sustainably exploit but also protect karst water resources by establishing protection zones accordingly. Malard et al. (2015) explain that only few generalizable methods for karst spring catchment delineation based on models have been proposed. Instead, delineations usually rely on classical hydrogeological methods such as assessing geology, topography, hydrology, water balance, elaborate tracer tests, and geophysical investigations. These methods usually are complex and costly, thus for many karst springs, exact catchment delineations are not available at all or at least contain some uncertainties. Where no information about the catchment is available at all, an approximate localization is advantageous as a first step towards an exact delineation since it facilitates the application of more elaborate methods like tracer test. There has already been an attempt by Longenecker et al. (2017) to semi-automatically derive approximate catchment boundaries by correlating karst spring discharge events with global precipitation measurement (GPM) gridded data (NASA, 2016). The authors achieved reasonable results with their method but also noticed that they could not replace conventional methods.

Anderson and Radić (2022) already applied gridded meteorological data to streamflow modeling in western Canada and used a coupled 2D-CNN-LSTM model to directly process spatially distributed input data. They showed that such models learn the relevant parts of the large-scale gridded input data for each local or regional streamflow automatically. We adapt and extend this approach to karst spring discharge modeling, however, purely based on CNNs by replacing the LSTM part with a 1D-CNN. Similar to the approach of Anderson and Radić (2022), in our proposed methodology, the 2D-CNN part learns the spatial features of the input data, while the 1D-CNN part extracts the temporal features, both necessary to simulate the spring discharge time series. With this combined 2D-1D-approach (for the sake of simplicity in the following only 2D-approach), we can now directly use gridded meteorological data to potentially overcome the common data availability problems when using climate station data for modeling. This approach further does no longer depend on a prior description of the catchment area, other than a very rough estimation of its approximate size to select the gridded data section large enough. Moreover, we investigate the potential of this approach for identifying the approximate catchment location based on a modified spatial input sensitivity analysis from Anderson and Radić (2022). To derive recharge areas based on rainfall-discharge event correlation, as previously done by Longenecker et al. (2017), requires (i) heterogeneous rainfall at catchment scale, (ii) precipitation data with sufficient spatial

resolution that capture this heterogeneity, and (iii) a karst system without too much dampening of the precipitation signals. These requirements hold for our proposed methodology as well, but a potential advantage of ANNs is their nonlinearity which may better capture the nonlinear relationships between rainfall and discharge.

We explore the applicability of our proposed deep learning approaches with spatially distributed input data in modeling karst spring discharge in three different study areas in Austria (Aubach spring), France (Lez spring), and Slovenia (Unica springs). All three associated karst areas are well studied, and for Austria and France, several modeling publications are available as benchmarks. Discharge of Lez spring in France was extensively studied in the past, including several ANN studies. Please refer to Kong A Siou et al. (2011) for an overview of older modeling studies at Lez spring with approaches other than ANN. We omit three newer ANN studies because they either do not focus on modeling discharge (Kong-A-Siou et al., 2015) or train models not on the complete annual cycle (Sep.-Aug. in this region) but on single flash-flood events (Darras et al., 2015; Darras et al., 2017). The other ANN studies all use classical MLPs or recurrent MLPs for discharge modeling, and we introduce them shortly in the following. Kong A Siou et al. (2011, 2012) and Kong-A-Siou et al., 2013 use precipitation from three or six gauges, respectively, and all use a similar but slightly varying data basis of 12 to 13 full annual cycles between 1988 and 2006. Testing period is either the single cycle 2002/2003 (Kong-A-Siou et al., 2013, 2012) or two cycles roughly in the same period (2002-2004) (Kong A Siou et al., 2011). Kong-A-Siou et al. (2014) uses data from 1987 to 2007, however, this time additionally including evapotranspiration and pumping from the Lez aquifer. For Aubach spring in Austria, no ANN studies exist, however, other modeling studies are available. Three studies (Chen and Goldscheider, 2014; Chen et al., 2017c, 2018) based on three successive and improved versions of a combined lumped parameter (SWMM) and distributed model, investigate and simulate three springs of this karst system simultaneously. They all achieved high performance in terms of NSE ($>$0.8), but none of them covered a complete annual cycle as contiguous test period. Additionally, they differ considerably in terms of their individual data basis for modeling (number and position of climate stations used as input data), as well as their testing periods. The shortest test set only had 40 days (in autumn), the longest (Chen et al., 2017c) used one year of data for model calibration and performed a split-sample test on the same data set. This makes a comparison of modeling results among these studies difficult. For the third spring (Slovenia), several earlier modeling studies are available (e.g., Kaufman et al., 2020; Kaufmann et al., 2016; Kovačič et al., 2020; Mayaud et al., 2019), even including ANNs (Sezen et al., 2019), but none of these directly modeled Unica springs discharge, but rather focused on other aspects like cave hydraulics or polje modeling. Besides existing studies, we compare the results of the 2D-model with own 1D-CNN models using climate station input data to assess the

usefulness and possible advantages of the direct use of spatially distributed input data. As spatially distributed inputs, we use either hourly ERA5-Land reanalysis data (Muñoz Sabater, 2019) or daily E-OBS data (Cornes et al., 2018), depending on the temporal resolution of spring discharge data. We selected these datasets among all openly accessible datasets (e.g., via Copernicus Climate Data Store) because of their available variable set and their spatial and temporal resolution. We introduce them in more detail in the following data section. Finally, we explore the potential of the 2D-approach for karst spring catchment localization by investigating the spatial input sensitivity of the trained CNN models.

2 Data and Study Areas

2.1 Overview

In this study, we investigate three different karst springs: Aubach spring in the Hochifen-Gottesacker area in Austria (Figure V.1a), springs of Unica river in Slovenia (Figure V.1b) and Lez spring in southern France (Figure V.1c). All springs show different characteristics regarding relevant system properties (e.g., catchment size, complexity of the hydrological system), environmental conditions (e.g., dominant climate, anthropogenic forcing) and data availability (see also Table V.A1). All areas are well studied and existing data was easily accessible. Further, several previous modeling approaches are available for comparison, except for the Slovenian spring.

2.2 Aubach Spring, Austria

Aubach spring is a major karst spring in the Hochifen-Gottesacker karst area in the northern Alps at the border between Germany and Austria. Southern border of the area is the Schwarzwasser valley, which geologically forms the contact zone between the Helvetic Säntis nappe in the north and sedimentary rocks of the Flysch zone in the south (Goldscheider, 2005). In the northern part the dominant karst formation is the Schrattenkalk formation, a cretaceous limestone with a thickness of about 100 m. This Schrattenkalk is structured in folds, which hydrogeologically form parallel sub-catchments (Figure V.1a) that contribute to different proportions to the several springs in the valley (Chen and Goldscheider, 2014; Goldscheider, 2005). In this study we focus on one large, non-permanent spring called Aubach spring (1080 m asl, discharge up to 10 m^3/s). The Hochifen-Gottesacker area is largely influenced by seasonal snow accumulation and melting in the elevated regions (>1,600 m asl), which is also clearly reflected in the discharge of Aubach spring by increased baseflow and

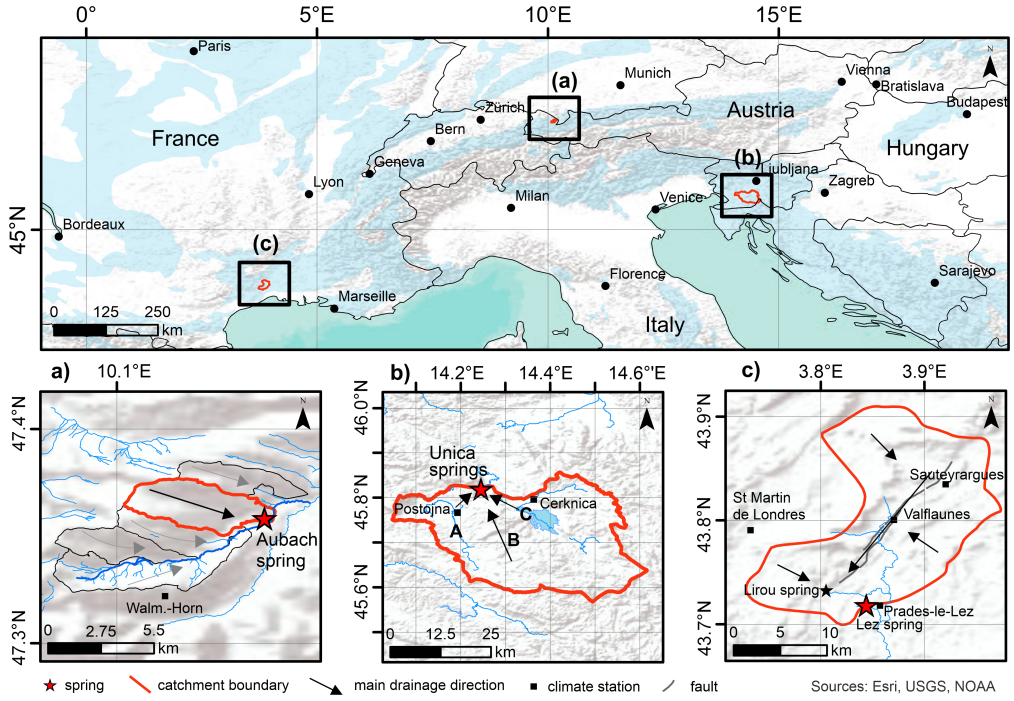


Figure V.1: Overview of all three study areas, the simulated springs (red star) and their catchments (red lines). Black squares indicate the locations of climate stations used for 1D-modeling (some are outside the shown maps), blue shadings in the upper map show karst areas based on WOKAM (Chen et al., 2017b) **a)** Hochifen-Gottesacker karst area and Aubach spring, black lines depict minor contributing sub-catchments; **b)** Unica river springs and Javorniki karst plateau (B); **c)** Lez spring catchment, Lirou overflow spring (black star) and major fault Corconne-Les Matelles (grey line);

diurnal snowmelt-induced variations, especially in the months of April to June. Earlier studies by Goldscheider (2005) and Chen and Goldscheider (2014) have identified one major catchment area of Aubach spring with approximately 9 km^2 (Figure V.1a), still, to smaller proportions upstream catchments can also contribute to Aubach spring discharge depending on the flow conditions. This applies also to the non-karstified Flysch area directly in the South (southernmost sub-catchment in Figure V.1a), where precipitation events are only relevant during low flow conditions. Then, the surface runoff from this area sinks into an upstream estavelle and contributes via an underground connection to the discharge of Aubach spring. During high flow conditions, the estavelle itself acts as an overflow spring and no contribution from surface runoff at Aubach spring occurs. Generally, the climate in the area can be described as cooltemperate and humid and the mean annual precipitation at the closest used climate station in this study (Walmendinger Horn) is about 2000 mm (2003-2019).

For this study we select Aubach spring because of the good data availability and use 8 years of hourly discharge data provided by the office of the federal state of Vorarlberg, division of water management. Further precipitation and temperature data from three surrounding

climate stations are available: Oberstdorf, Walmendinger Horn (shown in Figure V.1a) and Diedamskopf. Additionally, due to the high importance of snow in the area, we run a snowmelt routine as preprocessing of the meteorological input data as described in Chen et al. (2018). This routine is a slightly modified version (after Hock, 1999) of the HBV hydrological model snow routine (e.g., Bergström, 1975, 1995; Kollat et al., 2012; Seibert, 2000), which redistributes the precipitation time series in accordance with probable snow accumulation and snowmelt.

2.3 Unica Springs, Slovenia

The Unica springs (450 m asl) are located on the southern edge of a karst polje in SW Slovenia and are important from a biodiversity and water supply perspective. There are two permanent and several temporary springs that feed the Unica river. The joint discharge during 1989-2018 ranged from 1 to 90 m^3/s, while the mean discharge was 21 m^3/s (ARSO, 2020a). The springs are fed by three clearly distinguishable sub-catchments covering an area of about 820 km^2. The main recharge area is the highly karstified Javorniki plateau (up to 1,800 m asl; marked B on Figure V.1b), whose predominant lithology is Cretaceous rocks; mainly limestones, changing in places to dolomites and breccias. To a lesser extent, Jurassic and Palaeogene carbonate rocks are also present. The thickness of the unsaturated zone is estimated to be up to several hundred meters (Petrič et al., 2018, and references therein). To the east, a strike-slip fault zone controls the hydrology of the area, along which a chain of karst poljes developed (between 500 and 700 m asl; marked C on Figure V.1b). Upper Triassic dolomites predominate, changing to Jurassic limestones and dolomites in the south and west, forming aquifers with fracture porosity, which in places have very low to moderate permeability, and in some parts a superficial river network forms. As the karst poljes follow each other in a downward series, they are connected in a common hydrological system with transitions between surface and groundwater flows, and frequent flooding (Mayaud et al., 2019). In the West, the Pivka River Basin (between 500 and 700 m asl; marked A on Figure V.1b) consists of poorly permeable Eocene Flysch in the North, which conditions a surface river network. The southern part consists of Cretaceous and Jurassic carbonate rocks forming a shallow karst aquifer. Surface flow occurs during high water levels, receiving additional water from intermittent springs on the western foothills of the Javorniki plateau. The water flow of the sinking rivers in the subsurface from the regions A and C is of the channel flow type. We select the springs for this study because they drain a complex binary karst system of the so-called classical karst, they are well studied with long records of hydro-meteorological data and their hydrology is influenced by substantial snow accumulation and melting. The catchment belongs to the moderate continental climate and is mostly covered

with forests. For this study we use daily discharge data from the Unica-Hasberg gauging station (in the following called Unica) (ARSO, 2020a) and daily meteorological data from Postojna and Cerknica climate stations ranging from 1981 to 2018 (ARSO, 2020b). These climate stations (squares in Figure V.1b) are located on the western (Postojna) and eastern (Cerknica) part of the catchment, representing different climate regimes and are separated by the karst massif in between. For Postojna station the following variables are available: precipitation, temperature, potential evapotranspiration (PET), relative humidity, snow (S) and new snow (nS). For Cerknica station only P, S and nS exist. Average annual precipitation during 1989-2018 is about 1500 mm and on average 33 days of snow cover occur in Postojna (530 m asl) per year, while even longer snow cover is expected on the plateau.

2.4 Lez Spring, France

Our third study area is located 15 km north of Montpellier in southern France, within a large and complex karst system delimited by rivers and marly terrains. Eastern and western borders are the Vidourle and Hérault river valleys, northern and southern borders are piezometric limits. At larger scale, northern and southern boundaries are structural boundaries due to Cévennes and Montpellier faults, respectively. The dominant karst formations are Argovian to Kimmeridgian, and Berriasian massive limestones with 650 m to 1000 m thickness. Infiltration occurs mostly diffuse but also localized through fractures and sinkholes along the basin and through the major geologic fault of Corconne-Les Matelles in the northern part of the basin (indicated by a grey line in Figure V.1c).

The hydrogeological basin associated to the Lez spring has a size of about 240 km^2 (Figure V.1c), which is estimated on the basis of the hydrodynamic response to high discharge continuous pumping into the saturated zone of the aquifer (Thiéry and Bérard, 1983). However, the effective recharge catchment of the Lez spring, which corresponds to the extent of Jurassic limestone outcrops, has been estimated to be about 130 km^2 (Fleury et al., 2009; Jourde et al., 2014). The Lez karst aquifer is under anthropogenic pressure (i.e., aquifer exploitation for water supply) with pumping performed directly within the karst conduit. The discharge is measured at the spring pool and is regularly null during low water periods, when the pumping rate exceeds the natural spring discharge. Ecological water discharge towards the Lez river (160 L/s then 230 L/s after 2018) is ensured during such periods by a partial deviation of the pumped water to the river. Lirou spring (Figure V.1c) is the main of several overflow springs that activate during high flow periods (Jourde et al., 2014).

The Lez catchment is exposed to a Mediterranean climate, characterized by hot and dry summers, mild winters and wet autumns. Analyses by MeteoFrance show that on average

40% of the annual precipitation occurs between September and November with a high variability across years (Bicalho et al., 2012). The average annual rainfall rate for the 2008-2018 period is 904 mm.

For this study, we use nearly 10 years of daily discharge data provided by SNO KARST (Jourde et al., 2018; SNO KARST, 2021). The temperature data is from the Prades-le-Lez climate station; we use, however, an interpolated precipitation data series that is derived from a weighted average of four rainfall stations (Figure V.1c) (similar to Fleury et al., 2009; Mazzilli et al., 2011), three of them being located on the Lez catchment (Prades-le-Lez, Valflaunès, Sauteyrargues). The fourth station (Saint-Martin-de-Londres) is located few kilometers west of the catchment. Interpolation is in principle possible in this area due to the existing topography; at the same time, interpolation based on Thiessen-polygons (compare Appendix B) also allows compensation for data gaps at single stations. We decided to apply this preprocessing, because all but Saint-Martin-de-Londres climate station show such gaps from time to time, which explains the benefit from including within-catchment precipitation. We do not use pumping data as input in this study, because these were only available for a shorter period of time and such data would also not be available for a real forecast in the future (in contrast to weather and climate data).

2.5 Spatial Climate Data

Besides climate station data, we explored raster data from the E-OBS (Cornes et al., 2018), the ERA5-Land (Muñoz Sabater, 2019) and from the RADOLAN (DWD Climate Data Center (CDC), 2020) datasets as spatially distributed model inputs. E-OBS provides daily gridded meteorological data for Europe from 1950 to present, derived from in-situ observations, ERA5-Land provides hourly reanalysis data from 1981 to present. Both are available with a spatial resolution of $0.1° \times 0.1°$ (approx. $8\,\text{km} \times 11\,\text{km}$ for all study areas). Depending on the dataset, different sets of variables are available. In the case of E-OBS we initially provide our models with precipitation (P), mean, minimum and maximum temperature (T, Tmin, Tmax), relative humidity (rH) and surface shortwave downwelling radiation (Rad). For ERA5-Land, where a substantially larger set of variables is available, the following were used as initial inputs: total precipitation (P), 2m temperature (T), total evaporation (E), snowmelt (SMLT), snowfall (SF) and volumetric soil water of all four available layers (SWVL1: 0 - 7 cm, SWVL2: 7 - 28 cm, SWVL3: 28 - 100 cm, SWVL4: 100 - 289 cm). Relevant input variables from both datasets are later selected through Bayesian optimization (see section 3.3). The spatial extent of the input data is chosen very generously for each spring, so that between 6 and 8 additional cells are available as input data around the respective catchments. This prevents a predefinition of the area that needs to be identified as relevant as well as reduces

the influence of possible border effects due to the CNN approach using 3×3 filters (compare section 4.4). The resolution of ERA5-Land and E-OBS data corresponds to the grid cell size shown in the catchment plots in Figures V.1a-c, although each showing a slightly different absolute position of their grid center points. Depending on the temporal resolution of the available spring discharge measurements, we choose the spatial input data in accordance, thus E-OBS for Unica and Lez spring, ERA5-Land for Aubach spring.

Compared to the catchment size of Aubach spring (about $9\,km^2$), the spatial resolution (approx. $8\,km \times 11\,km$) of the gridded input data is extremely coarse. We therefore additionally explore a combination of ERA5-Land input variables (except P) with radar based precipitation data (RADOLAN) that offers a spatial resolution of 1×1 km^2 (DWD Climate Data Center (CDC), 2020). The higher resolved precipitation data from RADOLAN is thus augmented with climate variable values from ERA5-Land (for T, rH, etc.), which were downscaled and re-gridded to match the 1×1 km^2 RADOLAN grid. Compared to the ERA5-Land section around Aubach spring, for this additional analysis we reduce the spatial extent of the 2D-input data to save calculation time, but still considerably increase the total number of cells due to the higher resolution of the RADOLAN grid.

3 Methodology

3.1 Modeling Approach

In this study, we simulate karst spring discharge with deep learning models using meteorological input data. As proof of feasibility, we use meteorological data from surrounding climate stations as inputs to 1D-CNN models. However, data from such stations are often limited to precipitation and temperature, rarely more, as well as often exhibit data gaps, and limited record length or coarse sampling intervals. Also, the proximity to the catchment is often not sufficient, which especially in mountainous regions can introduce a distinct error in representing the true conditions within the catchment. This applies especially to variables with higher spatial variability such as precipitation.

Gridded meteorological data can be a solution to these issues, as they usually provide good temporal coverage and sampling intervals, a good spatial resolution as well as a large-scale availability (e.g., continental (E-OBS) or even global (ERA5-Land), see Bandhauer et al. (2021) for a comparison of both products). Further, especially reanalysis data include a larger variable set. When the catchment of the spring is unknown, it remains unclear which cells of the gridded data should be selected to best represent the climate conditions in the catchment, because the actual location of the spring is only a very rough indicator for

the location of the catchment. Based on our revised version of the approach of Anderson and Radić (2022), we demonstrate a solution by processing 2D-inputs and letting the model decide automatically, which parts of the input data are relevant to model the spring discharge.

3.2 Convolutional Neural Networks

Convolutional neural networks (LeCun et al., 2015) are widely applied in several domains such as object recognition (e.g., Cai et al., 2016), image classification (e.g., Li et al., 2014), and signal or natural language processing (e.g., Kiranyaz et al., 2019; Yin et al., 2017). The structure of most CNN models is based on the repetition of blocks that are made up of several layers, typically at least one convolutional layer followed by a pooling layer. The former matches the dimension of the input data (e.g., 2D for image alike data, 1D for sequences such as time series) and uses filters with a fixed size (receptive field) to produce feature maps of the input. The latter performs down-sampling of the produced feature maps and summarizes the features detected in the input. This decreases the total number of parameters of the model and makes it approximately invariant to small translations of the input (Goodfellow et al., 2016). A large variety of model structures based on such blocks, in combination with additional layers in between to prevent exploding gradients (e.g., batch normalization layers (Ioffe and Szegedy, 2015)) or model overfitting (e.g., dropout layers (Srivastava et al., 2014)) are possible; however CNNs usually end with one or several fully connected dense layers to produce a meaningful output.

From earlier studies (Jeannin et al., 2021; Wunsch et al., 2021) we know that 1D-CNNs are fast, reliable and excellently suited for modeling hydrogeological time series, such as groundwater levels or spring discharge. We have shown that they are faster compared to LSTMs, which are often the method of choice for time series modeling, and even outperform them or at least show similar performance (Wunsch et al., 2021). This is in agreement with the findings of (Van et al., 2020) in the domain of rainfall-runoff modeling. Based on these findings we choose CNNs for predicting karst spring discharge in this study and establish two different setups. One setup uses 1D-meteorological input data from surrounding climate stations and applies a 1D-CNN for forecasting. The second approach consistently uses a 1D-CNN to learn temporal features for discharge prediction, but combined with a time-distributed 2D-CNN to learn spatial features directly from gridded climate input data. Compared to the approach in Anderson and Radić (2022) we replace the LSTM by a 1D-CNN to make both setups methodologically consistent. Using CNNs in both setups further helps to assess the influence of using spatially distributed input data in terms of model accuracy, as we do not have to speculate if higher or lower performance might be due to the LSTM model rather than the input data. The general model structure of both setups is shown in Figure V.2.

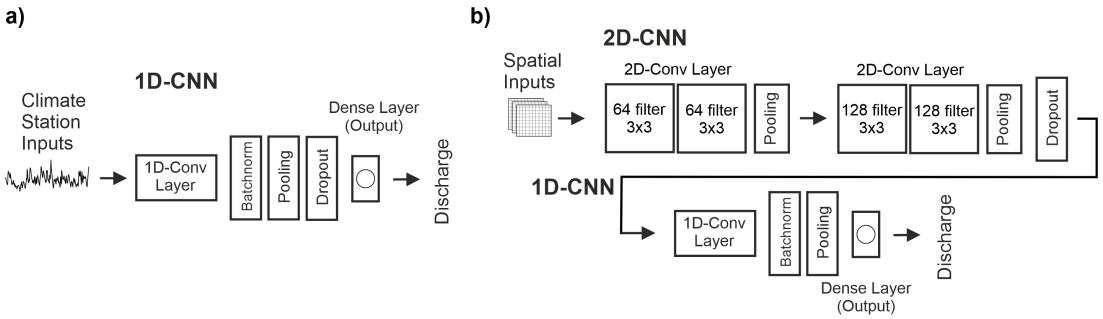


Figure V.2: Model structures applied for modeling karst spring discharge based on climate station data **a)** and gridded meteorological input data **b)**. Flatten layers are not displayed.

They basically use the same 1D-model except the position of the dropout layer. We use Bayesian hyperparameter optimization to select the 1D-filter number, batch-size and input sequence length of each model in both setups.

To reduce the dependency on the random initialization of the models, we use an ensemble with 10 members, each based on a different pseudo-random seed. Further, we implement Monte-Carlo dropout to estimate the model uncertainty from a distribution of 100 results for each of the ten realizations of each model in this study. We derived the 95% confidence interval from these 100 realizations by using 1.96 times the standard deviation of the resulting distribution for each time step. Each uncertainty was propagated while calculating the overall ensemble mean value for final evaluation in the test set. This uncertainty is shown as confidence interval for each of our simulation results in the following. We want to point out, that this uncertainty does not include other sources (such as input data uncertainty) but the random number dependency. All our models are implemented in Python 3.8 (van Rossum, 1995) and we use the following libraries and frameworks: Numpy (van der Walt et al., 2011), Pandas (McKinney, 2010; Reback et al., 2020), Scikit-Learn (Pedregosa et al., 2011), Unumpy (Lebigot, 2010–2020), Matplotlib (Hunter, 2007), BayesOpt (Nogueira, 2014), TensorFlow and its Keras API (Abadi et al., 2015; Chollet, 2015).

3.3 Model Calibration and Evaluation

We split the data for each site into four parts (Table V.1). The first part is used for training, the second part (validation) is simultaneously used to prevent overfitting via early stopping. The model's HPs are optimized according to its performance on the optimization set, while the last set is used as completely independent test set for final evaluation of the performance without data leakage from training or optimization. Training epoch number and early stopping patience are varied manually for each model at each test site. HPs for

the 1D-CNNs of both setups are optimized on the respective optimization set as stated above, maximizing the sum of NSE and R^2 (calculated as explained below). The number of optimization steps is varied manually for each model and is always a trade-off between accuracy and computational costs. In the case of many available input variables we treat input variable selection equally as a global optimization problem and use Bayesian optimization to simultaneously select a proper set of input variables and HPs. Thus, input optimization is used for each 2D-model, as ERA5-Land and E-OBS offer several different climate variables, as well as to the 1D-model of Unica springs, where the climate station records provide additional climate variables such as snow cover. For Lez spring and Aubach spring, only a smaller input variable set is available (mainly precipitation and temperature) and hence fully used. For all models we use an additional input (Tsin), which is a sine curve fitted to the temperature data. This variable can provide the model with noise-free information on seasonality and on the current position in the annual cycle (Kong-A-Siou et al., 2014). P is the only variable that is not optimized but fixed as input, because it has undoubtedly major influence on the discharge of a karst spring. The optimized HPs, information on some fixed HPs, and a summary of the number of parameters in each model, is given in Appendix Table V.D2. We calculate several

Table V.1: Data splitting schemes in years for all study areas (sample numbers in parentheses).

	Time Interval	Training	Validation	Optimization	Testing
Aubach spring	Hourly	2012-2017 (44,807)	2018 (8,760)	2019 (8,760)	2020 (7,320)
Unica spring	Daily	1981-2012 (11,687)	2013+2014 (730)	2015+2016 (731)	2017+2018 (730)
Lez spring	Daily	2008-2016 (2,629)	2017 (366)	2018 (365)	2019 (701)

metrics to evaluate the performance of our models: Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970), squared Pearson r, root mean squared error (RMSE), Bias as well as Kling-Gupta-Efficiency (KGE) (Gupta et al., 2009). For squared Pearson r we use the notation of the coefficient of determination (R^2), because we compare the linear fit between simulated and observed discharge, thus of a simple linear model, which makes them equal in this case.

3.4 Spatial Input Sensitivity and Catchment Localization

Anderson and Radić (2022) show in their study that combined 2D-CNN-LSTM models can learn to focus on specific areas of the spatially distributed input data and that these make physically sense. We modify this approach and transfer it to karst spring modeling, where we demonstrate that this approach is suited to approximate the location of karst catchments. We use the Gaussian spatial perturbation approach from Anderson and Radić (2022), which is similar to other input sensitivity algorithms such as occlusion (Zeiler and Fergus, 2014) or

RISE (Petsiuk et al., 2018), but in contrary to these methods takes into account the physical meaning of the absolute value of each variable at each pixel during the perturbation. We modify this approach so that only a single input channel (e.g., precipitation) is perturbed at a time for the sensitivity analysis. For details of this approach we refer to the original study. In short it works by perturbing spatial fractions of the input data by adding or subtracting a 2D-Gaussian curve from the input data at a certain location. Both the perturbed and unperturbed data are passed through the trained model to determine the resulting simulation error between them. In this way, after many iterations, heat maps are created that show how sensitive the trained model is to perturbations of certain areas of the input data. The considered input variables in our study show different properties in terms of spatial heterogeneity and variability. Temperature for example usually exhibits a distinct spatial autocorrelation, meaning that temperature information from outside the catchment area may be used to infer temperature within the catchment area. In contrast, precipitation is less spatially autocorrelated, meaning that precipitation information from outside the catchment area is less related to precipitation from inside the catchment area. Therefore, we hypothesize that the within-catchment precipitation fields will be most important for the model's prediction, and we will test this hypothesis by visually inspecting the sensitivity maps produced by the modified approach of Anderson and Radić (2022). Compared to the original approach by Anderson and Radić (2022), we therefore perturb only single channels at a time, instead of all channels at once, to separate the influence of each channel on the model output.

4 Results and Discussion

4.1 Aubach Spring

Figure V.3a shows the simulation results of the 1D-CNN model for the test period 2020, using only available climate station input data. Error measures indicate a high accuracy of the model simulation: NSE and R^2 values both are 0.74, KGE is 0.79. We observe that peaks in winter and spring are underestimated. The snowmelt period, clearly visible by increased baseflow and diurnal variations from April to June, is nicely fitted, as well as the following summer peaks. A short series of discharge peaks in the end of September/beginning of October is not captured. We assume that these were caused by small-scale precipitation events that are not represented in the data of the climate stations used as inputs. Interestingly, diurnal variations, which might be learned during the snowmelt period, are also visible in periods not influenced by snow (e.g., in August). From Chen et al. (2017c) we know the high relevance of snow in this area and by coupling the CNN model with a snow routine data preprocessing, we are able to further improve the model performance (Figure V.3b). We now

can achieve a fit with 0.77 for both NSE and R^2, KGE increases to 0.84. Our model is able to better fit the second largest peak of the whole dataset, which occurs in February, though, the peak is slightly overestimated, whereas other peaks still tend to be underestimated. The snowmelt period remains well simulated, but shows increasing deviations close the end of the period. The earlier noticed diurnal variations in summer and autumn, now are diminished, which is presumably an effect of the snowmelt preprocessing.

Please note that the 95% model uncertainty from random number dependency, estimated from 10 differently initialized models with a Monte-Carlo dropout distribution from 100 runs each (i.e., 1000 simulations in total), is very low for both modeling results (a+b) compared to the overall variability of the discharge. We assume the spatially limited input data to be the major source of error in the complete modeling procedure, because all climate stations are located outside of the catchment area and thus introduce distinct uncertainty about the true conditions within the catchment. Other modeling approaches (Chen and Goldscheider, 2014; Chen et al., 2017c, 2018) based on combined lumped parameter (SWMM) and distributed models, achieve similar or higher NSE values for the simulation of Aubach spring discharge (0.92, 0.83, 0.80 respectively). As mentioned, the results are, however, hardly comparable with each other and neither with this study. Reasons are (i) different input data (number and position of climate stations), (ii) different simulation periods, and (iii) very different test set lengths. One reason for the slightly lower performance of our model could be that none of the previous studies covered a complete annual cycle as contiguous test period, including high peaks in late winter and strong snowmelt influence in spring and early summer.

Figure V.3c shows the results of the 2D-modeling setup using (only) ERA5-land input data. Based on the described optimization procedure, the model uses the following inputs: P, T, E, SMLT, SWVL2 and SWVL4 (for a comparison of selected input variables with other study areas see also Table V.A1). The performance of the 2D-model is similar to that of the 1D-models, showing a NSE (0.76) and RMSE in-between both, a larger R^2 (0.8) but a lower KGE (0.71). This performance is still high considering that the major catchment is extremely small (about 9 km^2) compared to one ERA5-Land grid cell, and that a large grid section of 14×14 ERA5-Land cells ($1.4° \times 1.4°$) was used as input. We see that the major peak in February is slightly underestimated, as well as the beginning of the snowmelt period in April; however, the end of this period in May/June has improved now compared to (b). Both 1D-models are superior in estimating the peaks especially during summer, except the already mentioned peaks in September/October, which have improved using the 2D-input data. This supports the assumption that the climate stations do not represent these precipitation events, but the 2D-data does due to its spatial nature.

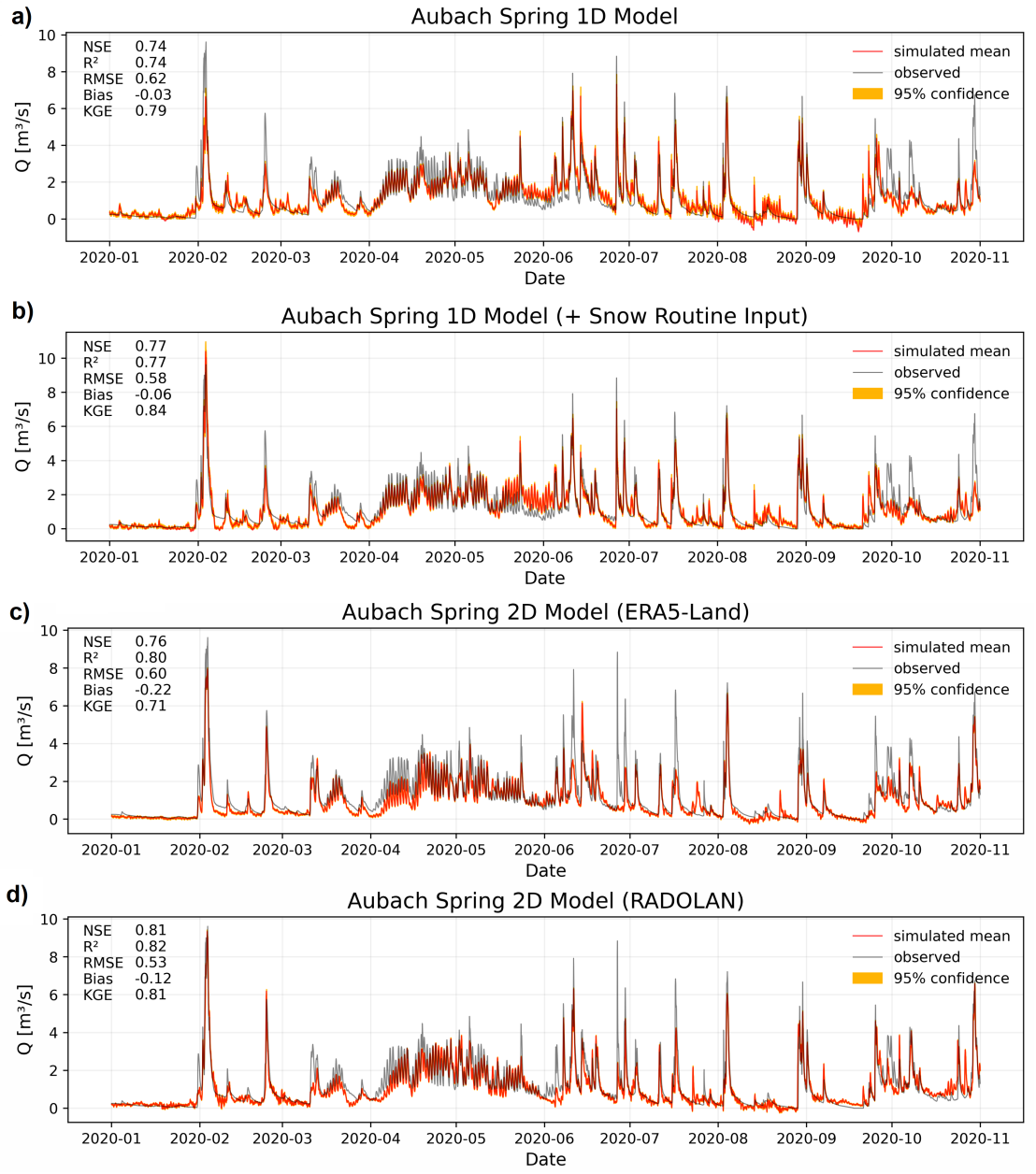


Figure V.3: Simulation results for the year 2020 at Aubach spring: **a)** 1D-model based on climate station inputs, **b)** 1D-model with additional snow routine preprocessing, **c)** 2D-model based on ERA5-Land gridded data and **d)** 2D-model with combination of ERA5-Land data and RADOLAN precipitation input.

To account for the small area of the catchment of Aubach spring, Figure V.3d shows the results of the 2D-input data, using the spatially higher resolved RADOLAN precipitation data in combination with downscaled ERA5-Land data. We have reduced the spatial extent of the 2D-input, but still have a reasonable buffer around the catchments and, compared to the former 2D-model, increase the grid cell number considerably (22×22 or $22^2\,km^2$). The optimized model uses P, T, Tsin, SMLT, SF, SWVL1/2/4 as inputs, thus omits E and SWVL3. This model shows the best performance of all four models by reaching a NSE of 0.81, R^2 of 0.82 and KGE of 0.81. Similar to the model in (c), the beginning of the snowmelt period in April remains slightly underestimated and compared to the 1D-models, the peaks in summer are less well fitted. Nevertheless, we generally see an accurate fit, especially the largest peak in February is well reproduced. Compared to the 1D-approach, the main source of uncertainty for both 2D-models should be the uncertainty of ERA5-Land variables. Their values originate from large grid cells in comparison to the catchment size, thus it is not clear how well they represent the true conditions on catchment scale. A more elaborated downscaling of ERA5 data or other high resolved climate data for a combination with RADOLAN precipitation data might be a promising approach for simulating small catchments like this one. Model uncertainty derived from random number effects and Monte Carlo dropout is (equally to the 1D-models) satisfyingly small. In total, we think that both the 1D and the 2D-approach for this catchment bear substantial shortcomings in terms of how well the input data represents the true conditions in the catchment, even though the simulation results are generally very accurate. On the one hand the climate stations represent the true observed climate, on the other hand this is true only for a very specific point, which is in this case outside the catchment, and embedded into a highly variable topography. The 2D-data have a too coarse spatial resolution compared to the size of the Aubach spring catchment and are themselves modeled (in the case of ERA5-Land). We therefore do not think that one approach is superior for this study area, but we can show that even in this case with relatively coarsely gridded input data compared to the catchment size, the 2D-approach offers a decent alternative in the case of missing climate station data.

4.2 Unica Springs

Figure V.4 summarizes the 1D- and 2D-model performance on the years 2017 and 2018 for Unica springs in Slovenia. The simulation of this quite large catchment area ($820\,km^2$) is based on the data of only two climate stations (Postojna and Cerknica). All available input variables from both stations except relative humidity from Postojna station and new snow from Cerknica station were used as inputs as selected by the Bayesian optimization model. The 1D-model shows good performance overall (NSE: 0.73, R^2: 0.79, KGE: 0.63), including

a response for all major discharge events. However, recession slopes especially in 2017 are underestimated substantially and the plateau shapes of the large peaks (e.g., January 2018) are not well captured, but rather simulated as multiple peaks. In general, many of the high flow events at this gauge have a quite long duration of days to even weeks resulting in such plateau-like shapes. This is due to the regular flooding of the polje. After the drainage areas of the polje are completely flooded, there is a progressive back-flooding and a steady rise in the water level, which makes it impossible to accurately monitor the flow conditions under these conditions. Therefore, during the plateau-like peaks, when we cannot observe the true flow; the peaks simulated by the ANN might be conceptually true, which is however not possible to evaluate. The peak in April 2018 is quite clearly underestimated, whereas the following low flow period (summer 2018) is slightly overestimated. Such overestimation might be due to small scale meteorological events that are detected by the climate stations, but do not well represent the conditions in the whole catchment area. It is also important to notice that between 2014 and 2018 substantial environmental changes occurred in the catchment (Kovačič et al., 2020). During this period a considerable amount of vegetation was destroyed by a series of large-scale forest disturbances. We expect the evapotranspiration changed due to changes in canopy interception, water use, and soil moisture. As a result, spring behavior has likely changed, because vegetation cover is an important element of the water balance and recharge events may have resulted in higher infiltration rates and more intense spring response, as well as more pronounced droughts. The effect of this environmental change on the model performance is hard to evaluate, because it is not part of the training data. However, the model was optimized and validated (early stopping) on a part of the period with environmental changes, which means that the model may infer some information on the changes from these periods (2014-2016). It is not expedient to exclude this change from model building, since this would require to shorten the time series to the period after 2018, thus loosing almost the complete data basis. Due to highly complex hydraulic behavior in this study area, which is for example related to already mentioned polje floodings and to a strongly variable water level in the system that varies also the catchment area, extracting the highly nonlinear precipitation-discharge is especially challenging. We generally observe less dynamics in terms of the number of flood pulse events compared to Aubach spring. In terms of intensity of hydrologic variability, discharge rates can vary by about two orders of magnitude. This is primarily due to the large size of the catchment area, the very high degree of karstification of the carbonate rocks, and the fact that the main spring may act as an overflow spring.

By using the 2D-input data from 18×21 E-OBS grid cells we were able to improve the model performance substantially (Figure V.4b), reaching now a NSE of 0.83, a R^2 of 0.84 and a KGE of 0.80. Model input variables are: P, Tmax, rH and Rad. We generally

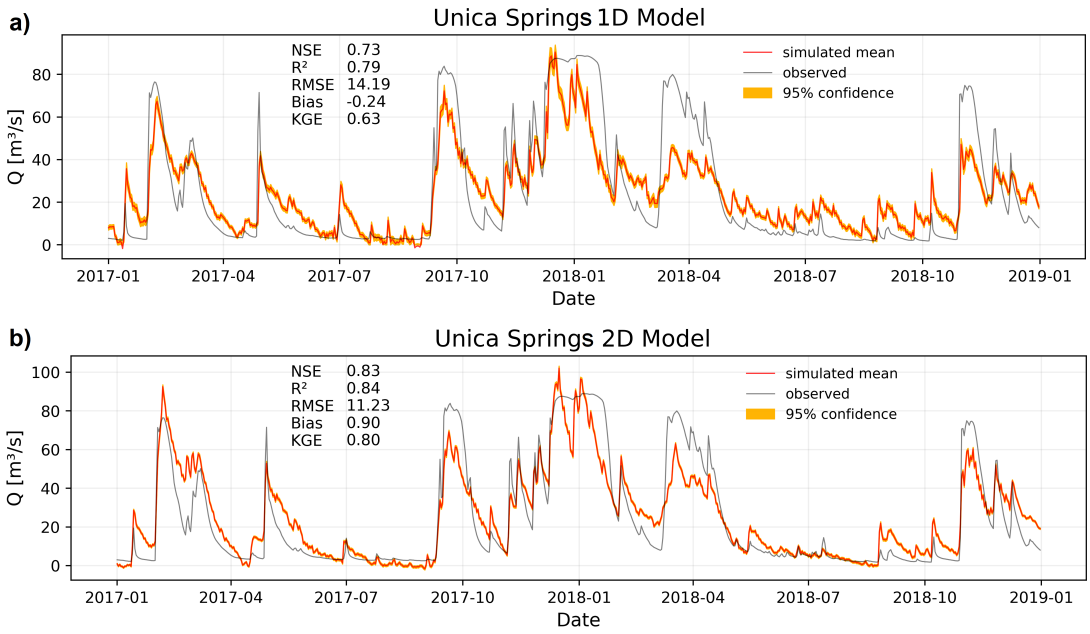


Figure V.4: Simulation results for 2017-2018 at Unica springs in Slovenia using **a)** climate station input data and **b)** E-OBS gridded data.

observe a similar shape of the simulation as for the 1D-model but with overall reduced errors. Still, the plateau shapes of some peaks are not well captured, but the same conceptual understanding as for the 1D-model seems to be learned, which means the model mainly simulates peaks instead of plateau-shaped high-flow events. The slope of the recessions are still generally underestimated, especially the simulation of low flow periods and minor discharge events improve clearly though. The improved results are plausible, because we can expect precipitation events to be represented more accurately in the gridded data that in the point data of only two climate stations, especially considering the size of the catchment. As for Aubach spring, both models show a comparably low model uncertainty based on random number variation and Monte-Carlo dropout, the model uncertainty of the 2D-simulation is even a bit lower than for the 1D-model. Again, we assume the spatially limited climate station data to be the main source of data uncertainty in the 1D-model, because meteorological stations are located on the western and eastern side of the karst massif. The massif itself represents the orographic barrier with different temperature and precipitation regimes that are certainly not captured by the considered meteorological stations. Concerning the 2D-data, the grid resolution is sufficiently high to adequately represent the climatic conditions in this large size catchment.

4.3 Lez Spring

Lez spring represents a third class of study area, as the catchment size (around 240 km^2) is somewhere in between the two others, the climate is Mediterranean and the spring runs dry for a considerable amount of time during the annual cycle due to a constant exploitation of the karst aquifer through pumping. Figure V.5 shows both the results for the 1D- (a) and the 2D-model (b). Despite comparably short training (daily data, starting in 2008) we observe a very high fit of the 1D-model above 0.86 for NSE, R^2 and KGE. As well the timing of the peaks, the absolute height of the peaks, as the dry periods are simulated accurately, except some deviations in early 2019.

For the 2D-model we use input data from 19×18 E-OBS grid cells and the Bayesian model selects only rH and Rad as inputs besides the fixed input P. Considering the high relevance of PET in the Mediterranean, it is a bit surprising that temperature, as a major driver of PET, is not selected (neither T, Tmin nor Tmax). However, relative humidity is also important to calculate PET (King et al., 2015) (e.g., low rH favors high evaporation) and the information on seasonality well encoded in a temperature time series, is presumably deducible from the radiation data (higher in summer than in winter) . The performance of the model is very good, but clearly lower compared to the 1D-model, showing NSE, R^2 and KGE between 0.75 and 0.78. Generally, the simulation is better in 2018 than in 2019, which is, however, also a tendency of the 1D-model. The model simulated some non-existent peaks in the dry sections, after all one of them (in Oct. 2018) clearly occurs also in the 1D-model's simulation. Presumably, the input data is accountable for the general performance differences between both modeling approaches. The climate stations, from which the interpolated precipitation time series is derived, are mainly located inside the catchment and additionally represent a good spatial coverage. Compared to both other study areas, the 1D-input data here best represents the climatic conditions within the catchment. Based on the lower performance of the 2D-model, we conclude that it seems to be harder to extract the relevant relationships between climate forcing and spring discharge from the gridded data. This may be related to a less favorable ratio of grid cell size to the catchment size, than for the Unica catchment for example. The model uncertainty based on initializations and derived from Monte-Carlo dropout again is small for both model setups, especially during dry periods.

The results of our models (1D-NSE: 0.87, 2D-NSE:0.75) can compete with the results from several earlier studies (NSE: 0.76-0.88 (Kong A Siou et al., 2011), NSE: 0.76-079 (Kong-A-Siou et al., 2014)), however, we do not beat the maximum performance reported by Kong A Siou et al. (2012) (NSE: 0.69-0.95) and Kong-A-Siou et al. (2013) (NSE: 0.96). Generally, all studies, including ours, achieve high performance and it is hard to conclude reasons for the superiority of one or other study, as several factors differ among them, such as model

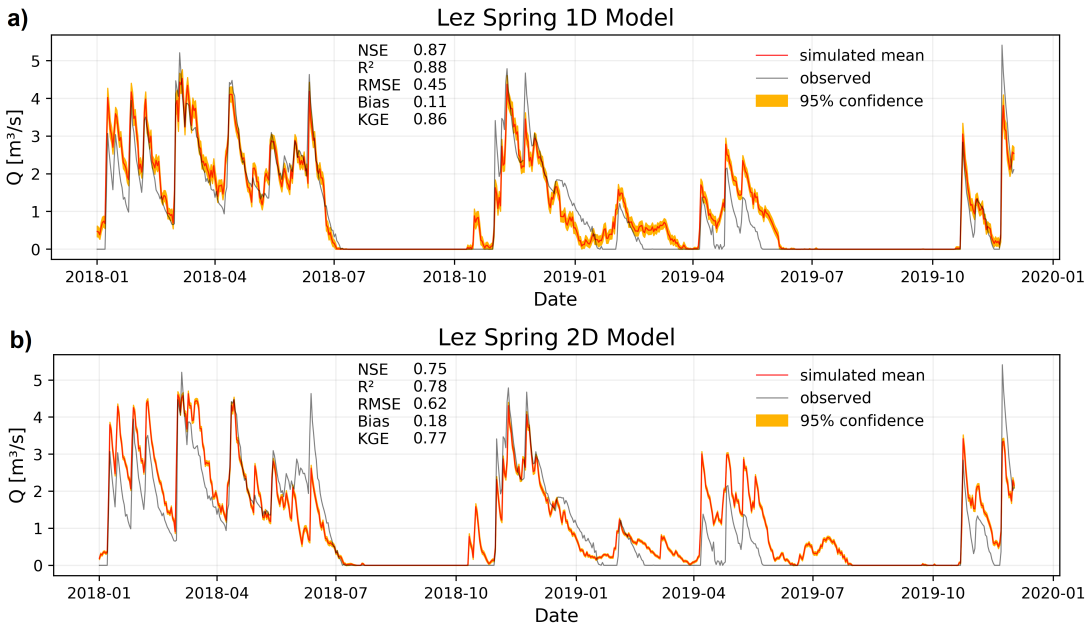


Figure V.5: Simulation results for 2018-2019 at Lez spring in France using **a)** climate station input data and **b)** E-OBS gridded data.

types, training and testing periods, or set of input variables. For our study, we chose not to include pumping data (as used in Kong-A-Siou et al. (2014)) due to the data availability reasons elaborated in section 2.4, as well as to be consistent in the 2D-modeling approach, which would need an update of the model structure due the 1D-time series character of the pumping data. The 2D-approach still shows very good performance in general, however, in comparison among all mentioned NSE values its performance is rather low. Nevertheless, we conclude if no climate station data would be available to apply a 1D-model, the 2D-approach still offers a reasonable substitute.

4.4 Spatial Input Sensitivity Results

The most important results of the spatial input sensitivity analysis from all catchments are shown in Figure V.6. In the case of Aubach spring modeled with ERA5-Land data (Figure V.6a), we can see that the catchment is smaller than one grid cell. Hence, despite the quite good discharge modeling, we see no clear spatial meaning of the precipitation channel heatmap. We also find a border effect with an almost uniform decrease in sensitivity toward the edges, which is an important reason to choose the spatial extent of the data large enough. This effect could be related to the size of the filter in the convolutional layer (3×3), as it sometimes only occurs in the one or two outermost pixels (e.g., Figure V.6c). In combination with zero-padding, which we apply to improve the informative value of the edges and to maintain the data size throughout the convolutions, this may result in such error halo, as also illustrated by Innamorati et al. (2020). Yet its origin remains unclear and not all heatmaps show this pattern (Figure V.6d), which questions the hypothesis of being a purely technical issue. For Aubach spring, precipitation shows only the fourth highest sensitivity (S) in terms of absolute values, while the second most sensitive variable is snowmelt (SMLT), which shows also the best spatial agreement with the catchment area. This is plausible insofar as the discharge for a large part of the time is dominated by snowmelt and to a lesser extent directly by precipitation. We conclude that even though the modeling results are satisfying, not much meaning can be extracted from the spatial sensitivity analysis for such a small catchment, given the existing spatial resolution of the gridded data. Please find heatmaps of all other variables in Appendix Figure V.C1. The combined approach of RADOLAN and ERA5-Land data (Figure V.6b) shows the heatmap in more detail in relation to the size of the catchment. We show only the precipitation heatmap, because it is the only variable with a native resolution of $1\,\text{km} \times 1\,\text{km}$ and we do not consider the spatial patterns of the remaining ERA5-Land-based variables to be meaningful to interpret. We observe that the most sensitive cells are identified close to the spring and at the border between the main catchment and the southern adjacent subcatchment. Due to the small scale of the spatial extent shown in Figure V.6b in relation to the spatial extent of precipitation events, the model is not able to sharply distinguish between precipitation inside and outside the catchment. This is presumably also related to the data, as precipitation is not directly measured, but estimated from radar signals and subsequently adjusted according to measured values from nearby climate stations. It remains unclear if precipitation is spatially resolved with sufficient accuracy in such alpine valleys on km-scale. No plausible reasoning exists for the two separate sensitive areas in the SW and NE corners, however, they are less sensitive then the center cells of the map.

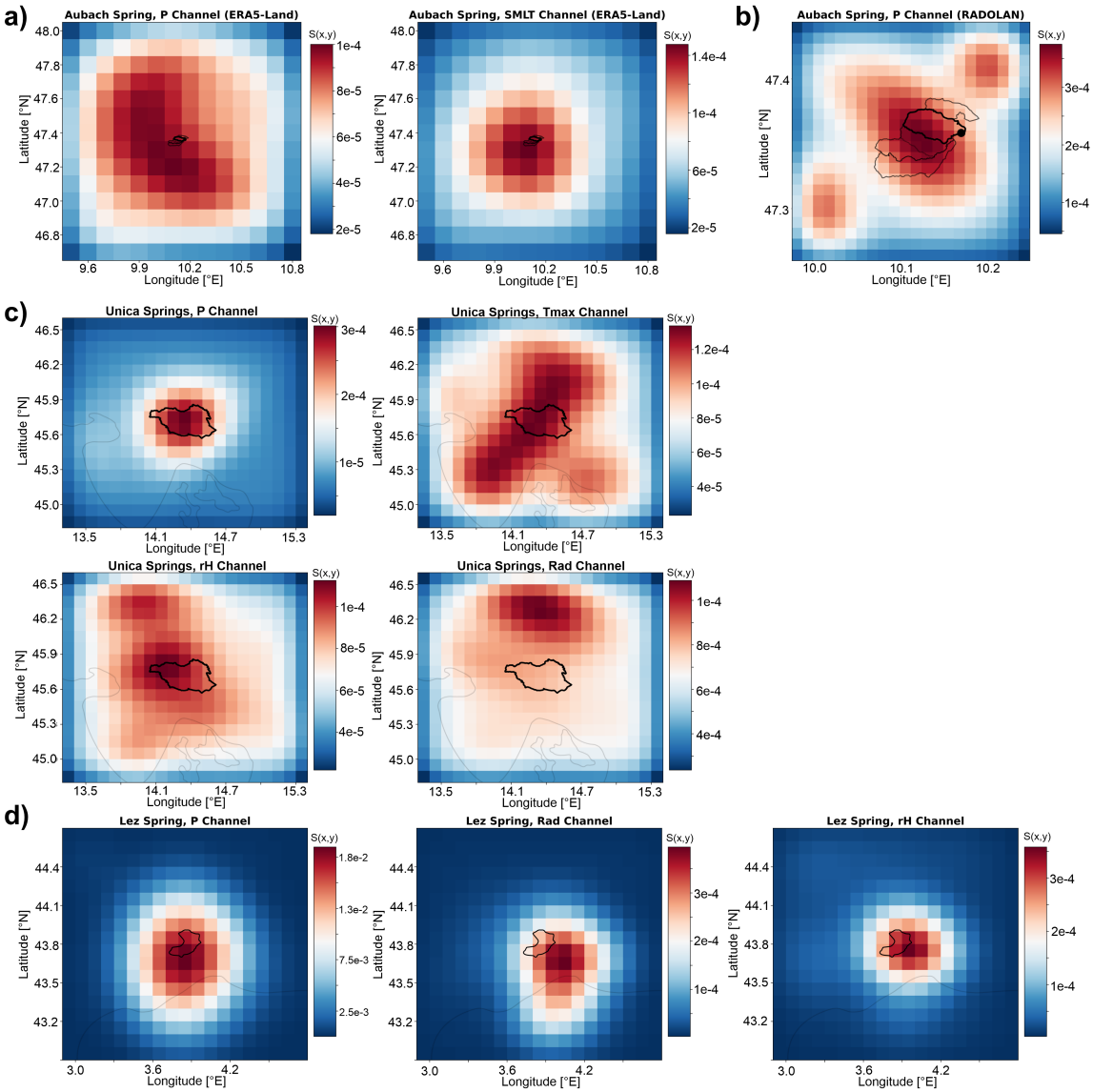


Figure V.6: Heatmaps of spatial input sensitivity for **a)** Aubach spring based on ERA5-Land gridded data, **b)** for Aubach spring based on RADOLAN precipitation data, **c)** Unica springs and **d)** Lez spring both based on E-OBS gridded data. In the case of **c)** and **d)**, light-grey lines indicate the coastlines for orientation.

Heatmaps of all four selected E-OBS variables at Unica catchment are shown in Figure V.6c. In accordance with our expectation for karst areas, we see the highest sensitivity for precipitation, which visually also identifies the catchment area very well. Especially Tmax and rH show high sensitivities on larger areas, however they are usually highly spatially autocorrelated and do not show a strong spatial heterogeneity like precipitation, which makes it plausible that the model learns from larger areas and does not concentrate strongly on the catchment itself. The model further identifies an area in the north as most sensitive for radiation.

Heatmaps of the 2D-Lez spring model are shown in Figure V.6d. In this area the model very strongly ignores large parts of the input data (dark blue, no visible border effects) and comparably sharply identifies the relevant area for the spring. This might be related to the higher spatial heterogeneity of precipitation in Mediterranean climate (Fresnay et al., 2012), which in this specific region has a special importance (severe flash floods known as Cévenol episodes (Kong A Siou et al., 2011)). Generally, we observe a slight south and east shift of the highest sensitivity compared to the catchment position. This might be related to the performance of the 2D-approach, which could not compete with the 1D-models. Maybe the model did not exactly learn the most relevant spatial features. The most sensitive variable is precipitation, while the rH channel shows the best spatial fit. We furthermore see that the size of the catchment is about the minimum size to produce meaningful heatmaps based on this given grid resolution, which corresponds also to our interpretation of the 2D-model performance shortcomings in comparison with the 1D-approach.

Given the spatial resolution of the used input data, the obtained heatmaps, and the simulation results of all three catchments, the Unica springs catchment seems to be most appropriate to further investigate the usefulness for catchment localization. It has the highest ratio of catchment size to data resolution and exhibits both generally high performance of the ANN models, and especially a considerably improved performance when using spatially distributed inputs compared to climate station input data. Thus, we used the Unica springs to conduct additional experiments to investigate the sensitivity of our approach to the absolute catchment location within the considered area of the input data. Figure V.7 shows the results of these experiments, where we shifted the 2D-input data boundaries in such a way that the catchment is located in one of the four corners or edges, leading to eight additional modeling results, named by the position of the catchment in the considered area of the input data. (e.g., *upleft*: catchment in the upper left corner). First of all, we find that all models successfully model the spring discharge curve and similarly learn the relevant grid cells of the considered input area, i.e., they are able to learn the approximate position of the catchment. The NSE values vary moderately between 0.80 and 0.85 among all models. The heatmaps of the precipitation input channel visually well identify the location of the catchment for each of the different considered areas of the input data. We find that regardless of the catchment's position within the considered areas of the input data, the resulting high-sensitive area in the P channel well indicates the true catchment position. For the heatmaps of the other input channels, we see that usually larger areas are identified as relevant and more variations between the models occur. Two things are particularly noticeable here. First, the identified sensitive input areas are generally slightly smaller for the *up** models, which is possibly related to the fact that the considered area of the input data is shifted towards the Mediterranean Sea, where no input data are available in the E-OBS dataset (compare the grey coastline).

These areas contain zeros or mean values and show no temporal variation that could be used to model the spring discharge. Second, the noticeably best performing model (*downleft*, NSE of 0.85) is the model with the least fraction of no-data cells (due to the Sea). Intuitively, we would not have expected the best performance here, but rather with the *upright* model, since there it is almost predetermined where the model has to learn. So the model seems to be able to use the larger amount of "useful data", even outside the catchment, to improve the overall performance. To possibly delineate a catchment from these results, a strategy has to be developed regarding the sensitivity contrast between the catchment and its surroundings. From our results we conclude that focusing on the precipitation channel is the most promising approach for potential catchment delineation. This makes, however, only sense if (i) precipitation is sufficiently heterogeneous at the scale of investigation, (ii) if conceptually spring discharge is mainly driven by precipitation (not snowmelt for example) and (iii) the gridded climate data is provided in a relatively high spatial resolution compared to the catchment size. Please find the precipitation channel heatmaps for Aubach spring and Lez spring in Appendix Figures V.C2 and V.C3.

In summary, we observe that the approach in its current form can produce meaningful heatmaps for at least roughly locating karst spring catchments. At least for the precipitation channel, we showed that the location of the catchment is successfully learned, regardless of the position within the considered area of the input data, if the ratio of catchment size to grid cell size is favorable (as for Unica springs). We notice that it generally works better the larger the catchment area, especially in relation to the grid cell size, but the absolute size of the catchment itself appears to be also important. For small catchments it seems harder to extract precise catchment locations, even if spatially finer-resolved data are available. This might be related to the fact that at small scales, even precipitation has a distinct spatial correlation, which can lead to higher sensitivity also in areas outside the catchment. However, one should keep in mind that these conclusions are only tendencies as we only investigated a small number of catchments. To develop a catchment delineation strategy, future investigations should analyze more catchments with adequate ratio of size to grid cell resolution, such as Unica catchment. Moreover, it can be expected that more and better gridded meteorological data products will be available in the future, which might lead to better results with the proposed methodology also for catchments with varying sizes.

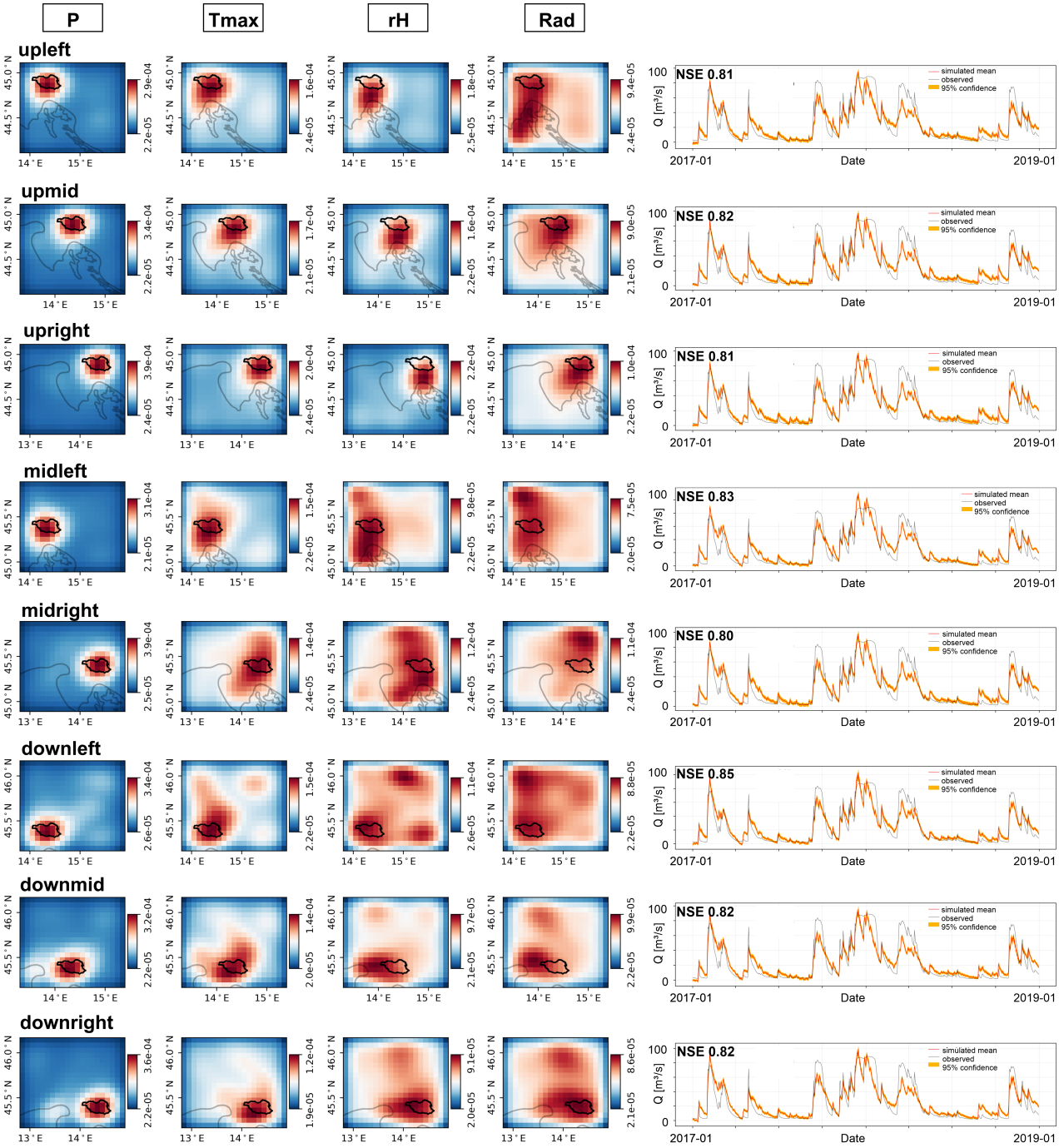


Figure V.7: Heatmaps of spatial input sensitivity for Unica springs catchment based on E-OBS gridded data. The considered area of the gridded input data is shifted to demonstrate the spatial learning capabilities of the models.

5 Conclusions

From the obtained insights, we can conclude that karst spring discharge can be predicted accurately with the presented 1D and 2D-approaches. Their performance competes with that of existing models in the three study areas. One main advantage compared to conventional modeling approaches in karst is that in order to obtain precise discharge simulations, far less prior knowledge of the system under consideration is required. Thus, using ANNs can generally reduce the amount of preliminary work that would be required to gain such sufficient system knowledge. We can further show that gridded climate data can provide an excellent substitute for non-existent or patchy climate station data. This does not require knowledge of the exact catchment area, which is a critical component, especially for karst springs. Rather, coupled 2D-1D-CNNs can be used to generate a first approximation of the catchment location. However, as it was shown, this approach still needs further development to more accurately localize the catchment, for example, by modifying the input sensitivity approach and by defining a routine to infer the catchment location from the sensitivity data, other than visual inspection. An important factor in achieving more accurate catchment localization is 2D-meteorological input data with a finer spatial resolution in relation to the catchment size because we found the approach to work best for the largest catchment. Additionally, a sufficient heterogeneity of precipitation in comparison to the catchment size is necessary, which, however, cannot be controlled but possibly limits the application in some karst areas. Given these developments and conditions, the approach's capabilities to delineate karst catchments should be further investigated, ideally including an evaluation against tracer tests and hydrogeological studies. In terms of accuracy, we do not find that one of the tested model setups (1D and 2D) is fundamentally superior. A key benefit of the 2D-approach, which uses spatially discretized input data, is the spatially and temporally complete nature of the data and the number of variables available for study. Furthermore, for many areas, the openly available 2D-climate data are easier accessible than climate station data, which still have to be collected from various different authorities, if accessible or existing at all. A weak spot of the 2D-approach is a substantially higher computational effort due to the large number of model parameters and the larger amount of data that has to be processed during training and optimization. In summary, gridded meteorological data is useful to overcome missing climate station data and to get a quite good idea of the spatial extent of larger catchments, given sufficiently small grid cell sizes.

Acknowledgments

The financial support of KIT through the German Federal Ministry of Education and Research (BMBF) and the European Commission through the Partnership for Research and Innovation in the Mediterranean Area (PRIMA) program under Horizon 2020 (KARMA project, grant agreement number 01DH19022A) is gratefully acknowledged. We thank the French Ministry of Higher Education and Research for the thesis scholarship of G. Cinkus as well as the European Commission and the Agence Nationale de la Recherche (ANR) for its support of HSM and UMR through the Partnership for Research and Innovation in the Mediterranean Area (PRIMA) program under Horizon 2020 (KARMA project, ANR-18-PRIM-0005). We further acknowledge financial support by the Slovenian Research Agency within the project Infiltration processes in forested karst aquifers under changing environment (No. J2-1743). The authors acknowledge support by the state of Baden-Württemberg through bwHPC. Muñoz Sabater (2019) was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store. The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains. We acknowledge the E-OBS dataset and the data providers in the ECA&D project (https://www.ecad.eu), data from MeteoFrance, DWD and the office of the federal state of Vorarlberg, division of water management. Lez spring discharge data were provided by the KARST observatory network (SNO KARST) initiative from the INSU/CNRS (FRANCE), which aims to strengthen knowledge-sharing and to promote cross-disciplinary research on karst systems.

Appendix

A Study Area Comparison Table

Table V.A1: Summary and comparison of different aspects of all three study areas.

	Aubach Spring	**Unica Springs**	**Lez Springs**
Country	Austria	Slovenia	France
Climate	cooltemperate and humid	moderate continental	mediterranean
Catchment Area [km^2]	9	820	240
Mean Precip. [mm/year] (Station, Period)	2000 (Walm.-Horn, 2003-2019)	1500 (1989-2018)	904 (2008-2018)
Spatially distributed input datasets	ERA5-Land, RADOLAN	E-OBS	E-OBS
Offered variables	P, T, Tsin, E, SMLT, SF, SWVL1-4	P, T, Tmin, Tmax, Tsin, rH, Rad	P, T, Tmin, Tmax, Tsin, rH, Rad
Selected variables	ERA5-Model: P, T, E, SMLT, SWVL2, 4 RADOLAN-Model: P, T, Tsin,SMLT SF, SWVL1, 2, 4	P, Tmax, rH, Rad	P, rH, Rad
Omitted variables	ERA5-Model: Tsin, SF, SWVL1, 3 RADOLAN-Model: E, SWVL3	T, Tmin, Tsin	T, Tmin, Tmax, Tsin

B Lez Catchment Precipitation Interpolation

The Thiessen's polygon interpolation method consists of calculating a weighted average of the precipitation data by allocating a contribution percentage to each meteorological station, based on its influence area on the catchment. These influence areas are calculated through geometric operations. First, we draw straight-line segments between each adjacent station, then we add the perpendicular bisectors of each segment, which will define the edges of the polygons. Each meteorological station thus corresponds to a particular polygon, for which the precipitation over the surface is assumed to be the same as the measured precipitation at the station.

The weighted average of the precipitation P_{wa} at each time step is calculated as follows:

$$P_{wa} = \frac{\sum_{i=1}^{n} A_i P_i}{A} \tag{V.1}$$

With n the number of meteorological stations, A_i the area (over the catchment) of the polygon corresponding to the i^{th} station, P_i the precipitation measured at the i^{th} station and A the area of the catchment.

C Heatmaps

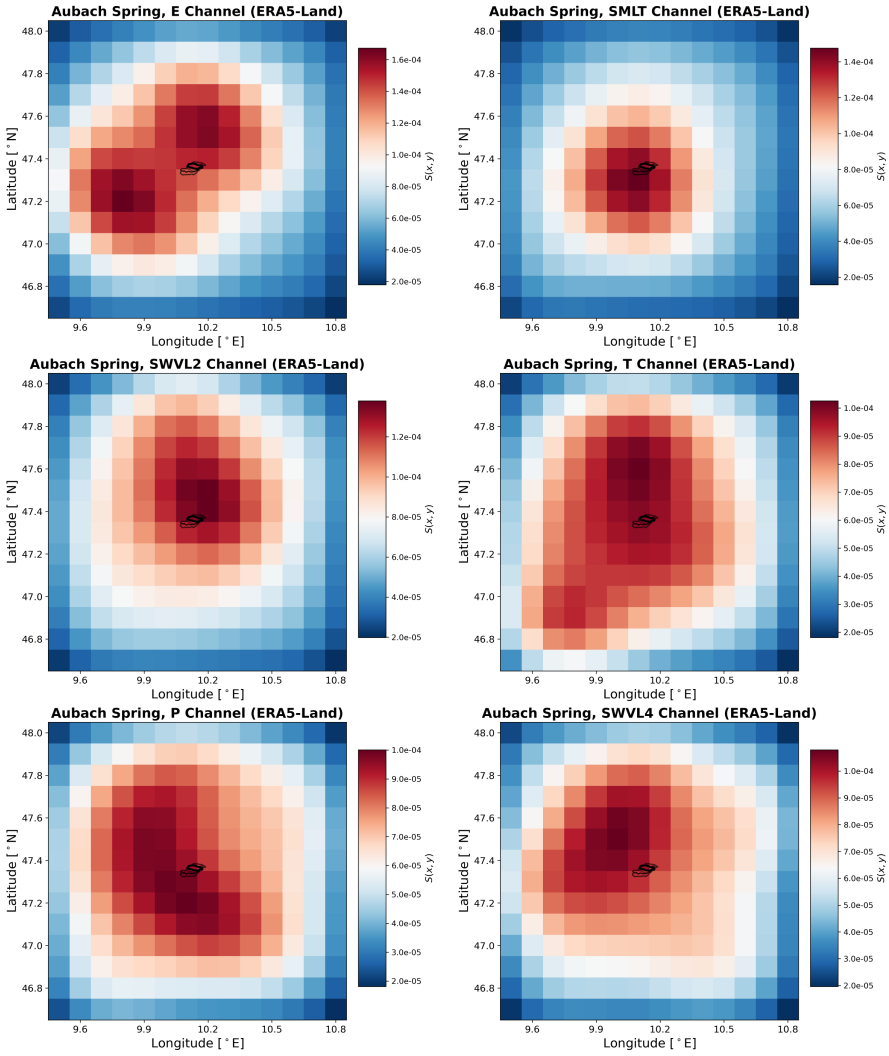


Figure V.C1: Spatial input sensitivity heatmaps for Aubach spring based on ERA5-Land gridded data.

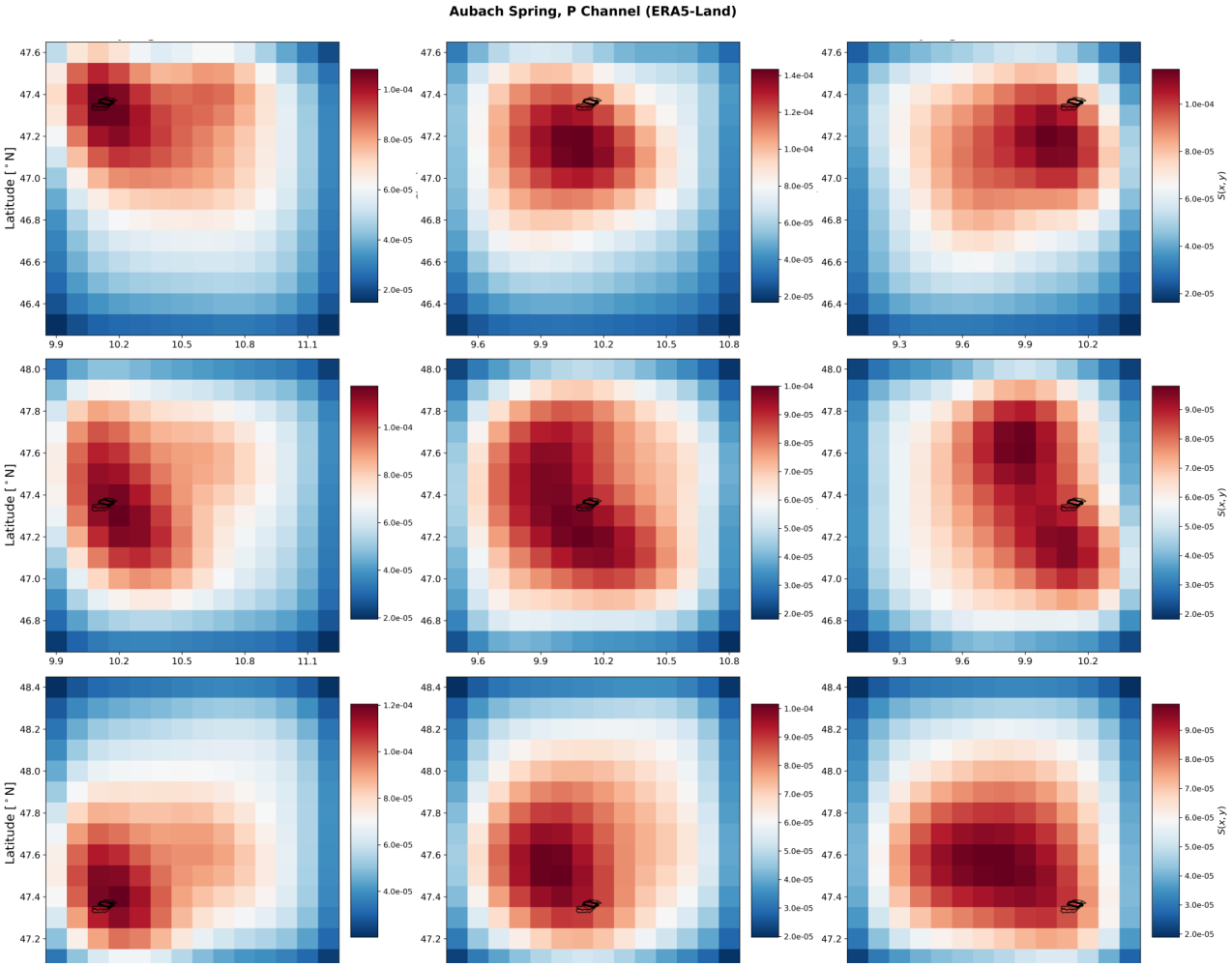


Figure V.C2: P-channel heatmaps based on ERA5-Land gridded data for Aubach spring with shifted area of the spatial input data in relation to the catchment position.

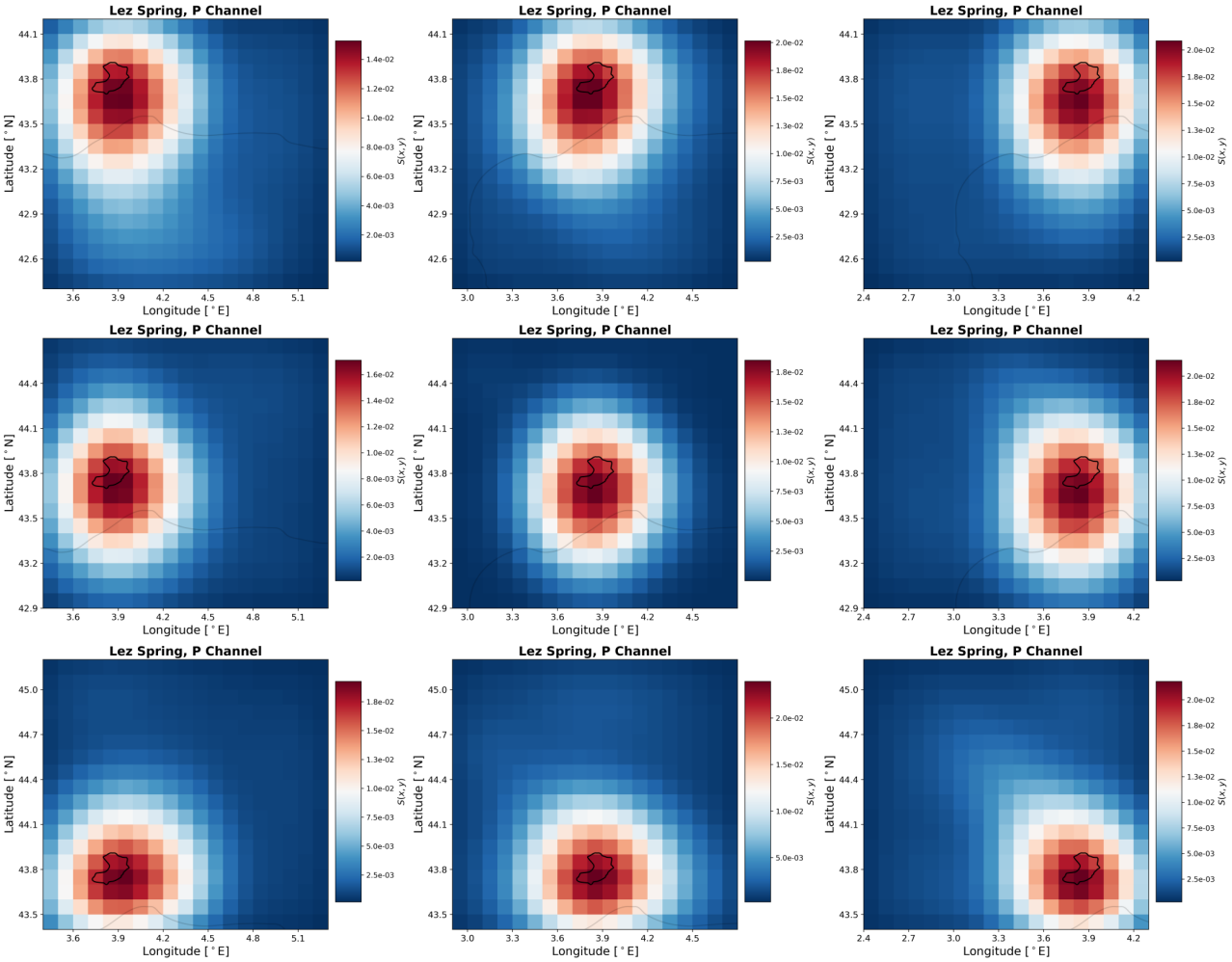


Figure V.C3: P-channel heatmaps based on E-OBS gridded data for Lez spring with shifted area of the spatial input data in relation to the catchment position.

D Model Overview

Table V.D2: Model parameter summary table.

	Aubach (ERA5)	**Aubach (RADOLAN)**	**Lez**	**Unica**
Optimized HP				
n (1DConv filter)	128	128	16	16
input seq. length	54 (hours)	162 (hours)	53 (days)	40 (days)
batch-size	64	256	32	32
Optimized Inputs	P (fixed)	P (fixed)	P (fixed)	P (fixed)
Yes	T, SMLT, E, SWVL2+4	T, Tsin, SMLT, SF, SWVL1+2+4	Tmax, rH, Rad	rH, Rad
No	Tsin, SF, SWVL1+3	E, SWVL3	T, Tmin, Tsin	T, Tmin, Tmax, Tsin
Other HPs				
inital learning rate	0.001	0.001	0.001	0.001
training epochs	100	100	100	100
early stopping patience	12	8	12	12
Model Summaries				
Total Parameters	708,353	1,502,849	358,977	384,017
Trainable Par.	708,097	1,502,593	358,945	383,985
Non-trainable Par.	256	256	32	32

Chapter VI

Synthesis and Outlook

1 Synthesis

ANNs pose manifold possibilities for modeling natural systems across various disciplines, including hydrogeology. This thesis explored several applications of ANNs to hydrogeological time series and successfully established approaches for hydrograph clustering, groundwater level forecasting on different time scales, and karst spring discharge prediction.

Chapter II, a study on the dynamics of groundwater level time series mainly designed to learn about their influencing factors, showed that homogeneous groups of groundwater hydrographs are not only interesting from a purely hydrogeological point of view. They are also helpful as preliminary work for groundwater modeling activities of any sort in a particular area. Knowledge about how a specific groundwater hydrograph represents the dynamics of a larger group proves itself highly valuable. Therefore, it is possible to select representative wells and model or predict them as surrogates for an entire group, which has a favorable implication for the required amount of work and computational power. Especially the latter, despite all progress in computing, still represents a hurdle for many applications and should not be neglected. The study aimed to provide a flexible framework with highly robust results based on ensemble approaches and, most importantly, an automated cluster number determination. Even though successful in these terms, the computational workload for datasets requiring a description with a larger feature number is undoubtedly a weak spot. Though reducing the potential workload of subsequent modeling actions, the method itself can be computationally quite expensive. Some guidance for dealing with these cases was provided alongside the study. In hindsight and relation to subsequent studies, a simple by-product of chapter II revealed itself to be a highly valuable outcome: the possibility to improve the data basis of individual wells based on cluster knowledge. The data of chapters III and IV were hence improved this way.

Chapter III is a state-of-the-art modeling approach comparison for groundwater level forecasting using ANNs with some surprising results. First, despite the focus of most related studies on popular DL approaches, the results showed that shallow network architectures such as NARX still deserve attention, especially since they achieved the most accurate results. Second, CNNs proved to be the most appropriate tool for modeling groundwater levels. They showed (i) slightly lower performance than NARX but considerably higher than LSTMs, (ii) the considerably lowest computational requirements among all models, and compared to NARX (iii) lower dependency on random initialization, and (iv) higher implementation flexibility. The experiments further showed that at least short-term forecasts are possible without any future input data. This finding is especially valuable for operational forecasting and respective short-term management decisions. Furthermore, the data requirements to build and train models that are capable of performing reasonable forecasts should be no high barrier for most applications in the groundwater sector, primarily because the presented approaches only focus on single site models, and first proper performance was achieved after ten years of weekly training data already. Generally, this study demonstrated the great potential of ANNs for forecasting GWL time series. From the hydrogeological point of view, modeling GWL with ANNs is comparably uncomplicated, as little domain knowledge and few input parameters are required, and the necessary gap-free GWL time series can often be easily prepared using the approach from chapter II.

Chapter IV demonstrated how to apply CNNs on long-term modeling tasks in the groundwater domain. These require special care in the model building process because the results of future periods cannot be validated. Additionally to validating the model performance in the past, both artificial climate scenarios to explore the behavior of the models in the extrapolating regime and an XAI approach proved to be helpful in enhancing trust in the model outcomes. Despite the innovative modeling approach of this study, the main contribution of this chapter is more on a non-technical level. This study clearly showed that even without considering important secondary factors such as anthropogenic withdrawals, we can expect large-scale decreases in groundwater levels in Germany until the end of the century. Overall, however, it becomes also clear that the changes for more optimistic scenarios, such as RCP2.6, are substantially smaller, indicating the importance of reducing global GHG emissions. It remains a future challenge to further refine and validate the results of this study using other modeling approaches.

Chapter V successfully transferred the existing modeling approaches from the groundwater domain to the related application of karst spring discharge modeling. However, the main contribution of this study is a specialized approach using spatially distributed input data, which showed that using openly available climate raster data can overcome insufficient climate station data availability within karst catchments. More importantly, ANNs can directly

process such data and independently learn the relevant spatial fraction. This study is only one of many recent examples where the classical division of DL methods into spatial learning (image-alike data) and sequential based learning (time series) fades, being also the typical scenario where DL stands out: exploiting spatial and temporal regularities in large amounts of data (Camps-Valls et al., 2021). This 2D-approach also introduced the exciting aspect of determining the location of catchment areas, which looks promising with some further development of the approach and an appropriate resolution of the spatial data.

To provide a final overview on the results of this thesis, the following paragraphs briefly answer the research questions formulated in chapter I:

RQ1: *How can we use unsupervised ANNs to group heterogeneous data sets of GW hydrographs based on their dynamics, and what can we learn from the resulting patterns?*

- Feature-based approaches are vital to make use of patchy real-world GW datasets.
- SOM+DS2L proved to be well suited for the intended purpose of clustering
- Similar GW dynamics are not only possible with spatial proximity.
- Influencing factors superimpose temporally and spatially and mostly are hard to disentangle. Nevertheless, some patterns are clearly dominated by distinct factors (e.g., surface water).
- The patterns provide valuable information on the representativeness of a single hydrograph's dynamic in relation to the region.

RQ2: *What are adequate model architectures to model and predict GWL time series, and what are their properties?*

- All models (NARX, CNN, LSTM) are capable of forecasting GWLs.
- NARX provide the most accurate predictions but are rather slow and sensitive to random initialization effects.
- CNNs are accurate, the fastest among all models, provide stable results, and are based on a flexible framework.
- LSTMs are outperformed in terms of accuracy by both other model types.

RQ3: *Is it possible to perform reasonable short-term predictions of GWLs with ANNs without any future input data?*

- Yes, for selected sites with reasonable performance for a 12-week forecast horizon.
- NARX provide the most accurate sequence predictions and almost keep up with CNNs in terms of computational speed.

RQ4: *What amount of data are necessary to build an ANN model for GWL prediction with reasonable performance?*

- The results are highly dataset-specific.
- For the given hydrographs, about ten years of weekly data were sufficient.
- It generally applies: the longer, the better; however, including recent periods with high relevance is more important than a training period as long as possible.

RQ5 *Can ANNs also be used to reasonably predict the long-term development of GWLs?*

- Yes, but only at sites with very high accuracy in validation periods in the past and given all relevant input variables for the future.
- Special care and additional analyses (e.g., XAI) are necessary to provide a trustworthy model.

RQ6: *How does the climate crisis influence the GWL development in Germany until the end of the century?*

- All investigated scenarios show decreasing GWL tendencies.
- GWL variability could potentially increase (found in RCP4.5 and 8.5).
- North and East Germany may be prone to stronger decreases than the South.
- RCP2.6 shows considerably less pronounced and less severe changes. The influence of stringently mitigating GHG emissions is clear.

RQ7: *How can state-of-the-art XAI techniques be used to increase trust in model decisions and to gain system understanding from ANNs models?*

- SHAP values provide valuable insight if models learn relations according to the existing conceptual understanding of relevant processes.
- Saliency maps are helpful to explore the spatial feature learning capability of 2D-CNN models.

RQ8: *How does a given routine for GWL modeling perform for predicting spring discharge in complex karst systems?*

- CNN models provide accurate karst spring discharge simulations and partly outperform existing approaches in the study areas.
- The approach is successful regardless of the specific karst system's properties (size, complexity, dominant climate, etc.).

RQ9: *Can ANNs learn the relevant fraction of spatially distributed input data automatically?*

- Yes, with high accuracy, regardless of the catchment size in all three study areas.
- The spatial fraction does not necessarily correspond to the exact catchment location (mainly depending on the input variable and the ratio of catchment size to grid cell size).
- Investigating the spatial sensitivity on precipitation input data might be helpful to locate karst catchments.

2 Outlook and Future Directions

DL research is making great strides across all disciplines, both improving or modifying existing model architectures and developing new model types whose potential for hydrogeology needs to be assessed. This does not only mean that the models in the presented approaches should be replaced with other/newer/better models, such as with transformers, which currently are at the foremost front of advances in DL research. Instead, completely new possibilities arise. For example, in the domain of rainfall-runoff modeling Kratzert et al. (2018) already showed in 2018 that one DL model trained on multiple basins on average outperforms specialized models that were specifically trained on a single site. Such more holistic modeling approaches offer attractive options, particularly for the studies from chapters III and IV, because a knowledge transfer takes place and the model now potentially generalizes across space, not only time. These approaches, however, also rely on large, currently mostly unavailable, data sets, which not only include GW data and relevant driving forces, but also descriptor variables (e.g., aquifer type, hydraulic conductivity, land use, vegetation, etc.) to distinguish between different sites. Especially in the case of climate impact modeling, future studies should also collect data from other climate zones (e.g., the Mediterranean) and apply only one model to all investigated sites. This strategy makes the model possibly more potent in estimating climate impacts (such as from dry periods) by generalizing from already seen conditions in the past but different areas. Given the relevance of karst areas for water supply, for example, in the alpine region, this aspect is also interesting for applications related to chapter V. Models could potentially generalize how systems react under the influence of glacier retreat in the future, using information from retreating glaciers in other areas in the past. Such investigation is not limited to a particular contiguous region such as the Alps in Europe but could potentially be extended to other alpine regions, such as in Asia or North America, to enlarge the data basis.

To model spatially closer and more distinct spatial relations between study sites (as in the URG from chapter III), spatio-temporal groundwater level forecasting is a promising option. Such approaches can simultaneously predict multiple points in space and time,

ideally considering the relation between neighboring points (or pixels, grid cells,...) instead of only providing unrelated predictions of such related points. Models for spatio-temporal forecasting either use uniformly gridded, thus image-alike data (e.g., using CNNs) or sensor networks (e.g., groundwater monitoring networks). Graph representations can account for such networks' spatially mostly irregular structure (basically nodes connected with vertices). Graph models using such graph structures usually apply temporal graph convolutions to learn from the similarity in the neighborhood. The basic idea is similar to CNNs in computer vision, assuming that neighboring pixels are typically related to each other somehow. However, graphs extend this idea to an irregular structure, contrary to CNNs that rely on regular, image-alike data with a uniform grid. Similar approaches already exist, for example, in the domain of traffic forecasting (e.g., Zhao et al., 2020; Zhu et al., 2021). Irregularly connected intersections, where traffic flow in one street often influences the roads nearby, can serve as an analogy to neighboring groundwater wells in the same aquifer. In reality, modeling groundwater monitoring networks should be more complex due to their additional dimension of depth and other difficulties, such as the fact that spatial proximity does not guarantee similar dynamics as shown in chapter II.

Understanding what AI models do and why they do it has a similar importance to developing new approaches to improve the modeling performance in general or to tackle unsolved modeling problems. Using XAI approaches, such as SHAP values in chapter IV or input saliency maps in chapter V are only two of many possibilities to do so. When applying DL models to a small sample size, we often end up with locally-fitted models, and we cannot expect them to learn universally-applicable physical laws necessarily (Shen et al., 2021). As in chapter IV, XAI is useful to check if a model nevertheless learns a representation of the input-output relationship that matches our physical understanding of the major processes. Such methods can considerably increase the trust in the model's decisions. The chance of DL models to learn more universal laws from large datasets is higher yet not assured; thus, XAI is helpful (even necessary) in this case, too. The difference to learning from small data sets is that we may use XAI to discover unknown relations from large data sets, which can potentially be translated into physical parameter hypotheses. As demonstrated in Tsai et al. (2020), conventional physically-based models can perform tests to investigate these hypotheses. Generally, XAI is gaining research interest because AI models get increasingly applied in critical sectors such as healthcare, and the pressure to justify model decisions amplifies (Gerlings et al., 2022). In hydrogeology, this might be less critical than in healthcare; nonetheless, explainability is vital to gain system understanding and learn about the underutilization of large data sets that researchers are unaware of in conventional approaches.

Besides understanding model decisions with XAI, we can also teach models to do the right things for the right reasons by introducing inductive biases through enforcing physical model constraints. Shen and Lawson (2021) argue that large potential lies in applications of such physics-informed or physics-constrained ML, especially for hydrogeology, where underground observations are limited. Additional information obtained by enforcing physical laws generally reduces the necessary amount of training data (Karniadakis et al., 2021; Shen and Lawson, 2021), but can also help in applications with "enough" data to address physical unknowns for improved generalization ability, predictive performance, and training speed (Karniadakis et al., 2021).

Table VI.1 summarizes the mentioned options for future research and highlights the currently limiting factors, that often prevent their application.

Table VI.1: Summary of the more advanced approaches mentioned, their respective relationship to the chapters, and currently existing problems that often hinder direct implementation.

Approach	**Relation to Chapter**	**Current Limitations**
holistic modeling: one model - many sites	III, IV, V	according data (esp. descriptor variables) are currently unavailable or hard to assemble
graph modeling: spatio-temporal forecasting	III	according data (esp. descriptor variables) are currently unavailable or hard to assemble; only few appropriate approaches exist yet.
XAI	III, IV, V	finding a suitable and compatible XAI approach; to whom should the explanation be useful?
physically informed models	II, III, IV, V	implementation of models is not yet straightforward; strong programming skills necessary

The focus of ANN users should not only be on improving model performance and building ever larger and more powerful models. Especially working in a field like hydrogeology that aims to understand and protect our environment, we should always consider the impact of what we do. This also applies to the algorithms that we apply to our data. DL works with large data, and processing such data in complex DL models on a global scale already needs a considerable amount of energy. This aspect also applies to this work, where especially the approaches from chapters II (SOM-Clustering) and V (2D-CNN models) potentially need considerable computing resources and a particular calculation time, depending on the specific problem. The field of sustainable AI is just in its infancy and is mostly a side aspect in current DL applications. However, awareness throughout the communities grows, and the goal should be not only to use AI for sustainability but also to use sustainable AI (van Wynsberghe, 2021).

This includes (i) the aspects of the hardware powering with sustainable energy and developing efficient hardware, (ii) developing and using efficient algorithms for training or applying models, or (iii) providing and exchanging pre-trained models so that each new task does not start from scratch. The latter aspect is also related to the aspect of embedding human knowledge in AI systems. Introducing knowledge (e.g., physical constraints as mentioned above) means that this knowledge does not have to be extracted from the data, which may help to reduce the associated energy consumption of such models (Vinuesa et al., 2020). We should also evaluate if DL is always the way to go. Using highly specialized models, both from the ML domain or other conventional approaches from a specific field of research may be as good as DL models (or better), however, circumventing the mentioned problems of DL. It is essential to raise awareness in all disciplines and follow sustainability principles in the future. Raising awareness also involves appropriate education that addresses these aspects and equips future AI practitioners with the necessary tools (s.a., programming skills) to apply algorithms efficiently or develop them themselves.

It will remain a challenge to keep up with ML research's current and future pace. On the one hand, advances in machine learning methods across disciplines are constantly yielding new methodical options; on the other hand, AI research is also being advanced explicitly in the hydrological sciences themselves (e.g., LSTMs on multiple timescales (Gauch et al., 2020)). The real benefit of DL lies in large sample sizes; one reason why relevant innovations and advances, aside from DL research itself, for water-related sciences mainly occur in rainfall-runoff modeling, as there are excellent (large-sample) datasets available. These are mostly missing for hydrogeology, even though lots of data are potentially available. Many of the ideas mentioned in the last paragraphs rely on appropriate data, which is usually the limiting factor when designing applications and new studies. The compilation of suitable hydrogeological data sets with many predictors and proper spatial and temporal resolution for Germany, Europe, and worldwide is, therefore, an essential task in hydrogeology in the coming years. Such datasets will provide the necessary conditions to keep pace with advances in DL research and to be able to test new approaches in hydrogeology as well. However, data alone is not sufficient. As argued by Shen et al. (2018), DL is still a niche skill in the hydrological sciences community, and including DL in the future hydrogeology curricula is an important task to profit from the great possibilities to come from this field.

Acknowledgments

Completing this dissertation has been quite a journey, and I am grateful to everyone who has accompanied me along the way.

First, I want to thank my supervisor Nico Goldscheider for giving me the freedom to pursue this topic in my doctorate and for the support throughout the years.

The biggest thanks go to Tanja Liesch because this thesis would not have been possible without her. Thank you for the constant motivation, encouragement, guidance, and support in all matters. Thank you for your (actual or virtual) door that was always open for professional discussions or private conversations. You know best how important your contribution to this dissertation is.

Thank you, Stefan Broda, for reviving AI research in the Hydrogeology Department, your support, fruitful discussions, and your efforts to realize AI projects, which have contributed substantially to securing my funding over the years.

Thank you, Anne Johannet, for the valuable time in France, where I learned a lot about ANN research. I would also like to thank the Graduate School for Climate and Environment (GRACE) that made my stay in France possible, as well as to visit the AGU in San Francisco.

I want to thank all colleagues for the good working atmosphere in the last years. Special thanks go to Marc Ohmer for the many discussions and climbing sessions. Also, thank you, Simon Frank, Nikolai Fahrmeier, Jonas Weis, Dominik Richter, Julian Xanke, and all other colleagues in the Hydrogeology and Engineering Geology Departments, with whom I have mostly worked off-topic but always enjoyed spending time.

Thank you, Sophia Kaiser, Alexandra Galisson, and David Seiler, for proofreading this thesis.

Finally, thanks to my family, who always had my back in my endeavors, and special thanks to Sophia, without whom, especially the last two years of home office, would have been incredibly tough.

Karlsruhe, February 2022 *Andreas Wunsch*

Author Contributions

Chapter II

A. Wunsch (AW) and T. Liesch (TL) conceptualized the study, all authors contributed to the methodology. AW developed the software code, performed formal analysis, investigation and validation of the results, and visualized the results. AW wrote the original paper draft, S. Broda and TL contributed to draft review and editing. TL supervised the work.

Chapter III

A. Wunsch (AW) and T. Liesch (TL) conceptualized the study, all authors contributed to the methodology. AW developed the software code, performed investigation and validation of the results, and visualized the results. AW wrote the original paper draft, S. Broda and TL contributed to draft review and editing. TL supervised the work.

Chapter IV

All authors contributed to conceptualization of this study. A. Wunsch (AW) and T. Liesch (TL) contributed to the methodology. AW further wrote the software code, performed validation, formal analysis, investigation, visualization and wrote the original draft. S. Broda performed data curation activities. All authors contributed to reviewing and editing the draft. TL and SB both supervised the work and were involved in project administration.

Chapter V

A. Wunsch (AW), T. Liesch (TL) and N. Goldscheider (NG) conceptualized the study, AW and TL developed the methodology and software code, and validated the results. AW performed the experiments, and investigated and visualized the results. G. Cinkus (GC) and Z. Chen performed formal analysis, N. Ravbar (NR) and GC contributed to data curation activities. AW wrote the original paper draft with contributions from GC and NR. All authors contributed to interpretation of the results, and review and editing of the paper draft. TL and NG supervised the work.

References

Abadi, Martın, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: p. 19. URL: https://www.tensorflow.org/.

Abrahart, Robert J. and Linda See (2000). "Comparing Neural Network and Autoregressive Moving Average Techniques for the Provision of Continuous River Flow Forecasts in Two Contrasting Catchments". In: *Hydrological Processes* 14.11-12, pp. 2157–2172. ISSN: 1099-1085. DOI: 10.1002/1099-1085(20000815/30)14:11/12<2157::aid-hyp57>3.0.co;2-s.

Adamowski, Jan and Hiu Fung Chan (2011). "A Wavelet Neural Network Conjunction Model for Groundwater Level Forecasting". In: *Journal of Hydrology* 407.1-4, pp. 28–40. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2011.06.013.

Afzaal, Hassan, Aitazaz A. Farooque, Farhat Abbas, Bishnu Acharya, and Travis Esau (2020). "Groundwater Estimation from Major Physical Hydrology Components Using Artificial Neural Networks and Deep Learning". In: *Water* 12.1 (1), p. 5. DOI: 10.3390/w12010005.

Alley, William M., Richard W. Healy, James W. LaBaugh, and Thomas E. Reilly (2002). "Flow and Storage in Groundwater Systems". In: *science* 296.5575, pp. 1985–1990. DOI: 10.1126/science.1067123.

Alsumaiei, Abdullah A. (2020). "A Nonlinear Autoregressive Modeling Approach for Forecasting Groundwater Level Fluctuation in Urban Aquifers". In: *Water* 12.3 (3), p. 820. DOI: 10.3390/w12030820.

Anderson, Sam and Valentina Radić (2022). "Evaluation and Interpretation of Convolutional Long Short-Term Memory Networks for Regional Hydrological Modelling". In: *Hydrology and Earth System Sciences* 26.3, pp. 795–825. ISSN: 1027-5606. DOI: 10.5194/hess-26-795-2022.

Arnold, Taylor Baillie and Lauren Craig Tilton (2019). "Depth in Deep Learning : Knowledgeable , Layered , and Impenetrable". In.

ARSO (2020a). *Slovenian Environment Agency. Archive of Hydrological Data.* URL: http://vode.arso.gov.si/hidarhiv/ (visited on 12/05/2020).

– (2020b). *Slovenian Environment Agency. Archive of Meteorological Data.* URL: http://www.meteo.si (visited on 12/05/2020).

Baker, David B., R. Peter Richards, Timothy T. Loftus, and Jack W. Kramer (2004). "A New Flashiness Index: Characteristics and Applications to Midwestern Rivers and Streams". In: *Journal of the American Water Resources Association* 40.2, pp. 503–522. ISSN: 1093-474X, 1752-1688. DOI: 10.1111/j.1752-1688.2004.tb01046.x.

Balakrishnan, P. V., Martha C. Cooper, Varghese S. Jacob, and Phillip A. Lewis (1994). "A Study of the Classification Capabilities of Neural Networks Using Unsupervised Learning: A Comparison withK-means Clustering". In: *Psychometrika* 59.4, pp. 509–525. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/bf02294390.

Balugani, E., M.W. Lubczynski, L. Reyes-Acosta, C. van der Tol, A.P. Francés, and K. Metselaar (2017). "Groundwater and Unsaturated Zone Evaporation and Transpiration in a Semi-Arid Open

Woodland". In: *Journal of Hydrology* 547, pp. 54–66. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2017.01.042.

Bandhauer, Moritz, Francesco Isotta, Mónika Lakatos, Cristian Lussana, Line Bå serud, Beatrix Izsák, Olivér Szentes, Ole Einar Tveito, and Christoph Frei (2021). "Evaluation of Daily Precipitation Analyses in E-OBS (V19.0e) and ERA5 by Comparison to Regional High-Resolution Datasets in European Regions". In: *International Journal of Climatology*, pp. 1–12. ISSN: 1097-0088. DOI: 10.1002/joc.7269.

Barthel, Roland, Tim G. Reichenau, Tatjana Krimly, Stephan Dabbert, Karl Schneider, and Wolfram Mauser (2012). "Integrated Modeling of Global Change Impacts on Agriculture and Groundwater Resources". In: *Water Resour Manage* 26.7, pp. 1929–1951. ISSN: 1573-1650. DOI: 10.1007/s11269-012-0001-9.

Battaglia, Peter W., Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu (2018). *Relational Inductive Biases, Deep Learning, and Graph Networks*. arXiv: 1806.01261 [cs, stat]. URL: http://arxiv.org/abs/1806.01261 (visited on 12/10/2021).

Beale, H. M., M. T. Hagan, and H. B. Demuth (2016). *Neural Network Toolbox™ User's Guide: Revised for Version 9.1 (Release 2016b)*. In collab. with Inc. The MathWorks. URL: https://de.mathworks.com/help/releases/R2016b/nnet/index.html.

Bengio, Y., P. Simard, and P. Frasconi (1994). "Learning Long-Term Dependencies with Gradient Descent Is Difficult". In: *IEEE Trans. Neural Netw.* 5.2, pp. 157–166. ISSN: 1045-9227, 1941-0093. DOI: 10.1109/72.279181.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2014). *Representation Learning: A Review and New Perspectives*. arXiv: 1206.5538 [cs]. URL: http://arxiv.org/abs/1206.5538 (visited on 02/03/2022).

Bergström, Sten (1975). "The Development of a Snow Routine for the HBV-2 Model". In: *Hydrology Research* 6.2, pp. 73–92. ISSN: 0029-1277. DOI: 10/gkcqz5.

– (1995). "The HBV Model". In: *Computer Models of Watershed Hydrology*. Ed. by V. P. Singh. Colorado, USA: Water Resources Publications, pp. 443–476. ISBN: 0-918334-91-8. URL: https://www.cabdirect.org/cabdirect/abstract/19961904773 (visited on 06/03/2021).

BGR (2019). *Mean Annual Groundwater Recharge of Germany 1:1,000,000 (GWN1000)*. URL: https://www.bgr.bund.de/had (visited on 10/21/2019).

Bicalho, C. C., C. Batiot-Guilhe, J. L. Seidel, S. Van Exter, and H. Jourde (2012). "Hydrodynamical Changes and Their Consequences on Groundwater Hydrochemistry Induced by Three Decades of Intense Exploitation in a Mediterranean Karst System". In: *Environ Earth Sci* 65.8, pp. 2311–2319. ISSN: 1866-6280, 1866-6299. DOI: 10.1007/s12665-011-1384-2.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer. 738 pp. ISBN: 978-0-387-31073-2.

Bouwer, Herman (2002). "Artificial Recharge of Groundwater: Hydrogeology and Engineering". In: *Hydrogeology Journal* 10.1, pp. 121–142. ISSN: 1431-2174, 1435-0157. DOI: 10.1007/s10040-001-0182-4.

Bowes, Benjamin D., Jeffrey M. Sadler, Mohamed M. Morsy, Madhur Behl, and Jonathan L. Goodall (2019). "Forecasting Groundwater Table in a Flood Prone Coastal City with Long Short-term Memory and Recurrent Neural Networks". In: *Water* 11.5 (5), p. 1098. ISSN: 2073-4441. DOI: 10.3390/w11051098.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

Brienen, S., A. Walter, C. Brendel, C. Fleischer, A. Ganske, M. Haller, M. Helms, S. Höpp, C. Jensen, K. Jochumsen, J. Möller, S. Krähenmann, E. Nilson, M. Rauthe, C. Razafimaharo, E. Rudolph, H. Rybka, N. Schade, K. Stanley, and S. Brienen (2020). "Klimawandelbedingte Änderungen in

Atmosphäre und Hydrosphäre: Schlussbericht des Schwerpunktthemas Szenarienbildung (SP-101) im Themenfeld 1 des BMVI-Expertennetzwerks". In: in collab. with BSH, BfG, and BAW. DOI: 10.5675/expnbs2020.2020.02.

Cabanes, Guénaël, Younès Bennani, and Dominique Fresneau (2012). "Enriched Topological Learning for Cluster Detection and Visualization". In: *Neural Networks*. Selected Papers from IJCNN 2011 32, pp. 186–195. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2012.02.019.

Cai, Zhaowei, Quanfu Fan, Rogerio S. Feris, and Nuno Vasconcelos (2016). "A Unified Multi-Scale Deep Convolutional Neural Network for Fast Object Detection". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, pp. 354–370. ISBN: 978-3-319-46493-0.

Caiado, Jorge, Elizabeth Ann Maharaj, and Pierpaolo D'urso (2015). "Time-Series Clustering". In: *Handbook of Cluster Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton London New York, pp. 241–264. ISBN: 978-1-4665-5189-3 978-1-4665-5188-6.

Callaway, Ewen (2020). "'It Will Change Everything': DeepMind's AI Makes Gigantic Leap in Solving Protein Structures". In: *Nature* 588.7837 (7837), pp. 203–204. DOI: 10/fk8q.

Camps-Valls, Gustau, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein, eds. (2021). *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*. 1st ed. Wiley. ISBN: 978-1-119-64614-3 978-1-119-64618-1. DOI: 10.1002/9781119646181.

Chang, Fi-John, Li-Chiu Chang, Chien-Wei Huang, and I-Feng Kao (2016). "Prediction of Monthly Regional Groundwater Levels through Hybrid Soft-Computing Techniques". In: *Journal of Hydrology* 541, pp. 965–976. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2016.08.006.

Chang, Li-Chiu, Hung-Yu Shen, and Fi-John Chang (2014). "Regional Flood Inundation Nowcast Using Hybrid SOM and Dynamic Neural Networks". In: *Journal of Hydrology* 519, pp. 476–489. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2014.07.036.

Chen, Lu-Hsien, Ching-Tien Chen, and Yan-Gu Pan (2010a). "Groundwater Level Prediction Using SOM-RBFN Multisite Model". In: *Journal of Hydrologic Engineering* 15.8, pp. 624–631. DOI: 10.1061/(asce)he.1943-5584.0000218.

Chen, Qian, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen (2017a). "Enhanced LSTM for Natural Language Inference". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1657–1668. DOI: 10.18653/v1/P17-1152. arXiv: 1609.06038.

Chen, Yiheng, Bing Qin, Ting Liu, Yuanchao Liu, and Sheng Li (2010b). "The Comparison of SOM and K-means for Text Clustering". In: *CIS* 3.2, p268. ISSN: 1913-8997, 1913-8989. DOI: 10.5539/cis.v3n2p268.

Chen, Yitian, Yanfei Kang, Yixiong Chen, and Zizhuo Wang (2020). "Probabilistic Forecasting with Temporal Convolutional Neural Network". In: *Neurocomputing* 399, pp. 491–501. ISSN: 09252312. DOI: 10.1016/j.neucom.2020.03.011.

Chen, Zhao, Augusto S. Auler, Michel Bakalowicz, David Drew, Franziska Griger, Jens Hartmann, Guanghui Jiang, Nils Moosdorf, Andrea Richts, Zoran Stevanovic, George Veni, and Nico Goldscheider (2017b). "The World Karst Aquifer Mapping Project: Concept, Mapping Procedure and Map of Europe". In: *Hydrogeol J* 25.3, pp. 771–785. ISSN: 1435-0157. DOI: 10/f98h6g.

Chen, Zhao and Nico Goldscheider (2014). "Modeling Spatially and Temporally Varied Hydraulic Behavior of a Folded Karst System with Dominant Conduit Drainage at Catchment Scale, Hochifen-Gottesacker, Alps". In: *Journal of Hydrology* 514, pp. 41–52. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2014.04.005.

Chen, Zhao, Andreas Hartmann, and Nico Goldscheider (2017c). "A New Approach to Evaluate Spatiotemporal Dynamics of Controlling Parameters in Distributed Environmental Models". In: *Environmental Modelling & Software* 87, pp. 1–16. ISSN: 13648152. DOI: 10.1016/j.envsoft.2016.10.005.

Chen, Zhao, Andreas Hartmann, Thorsten Wagener, and Nico Goldscheider (2018). "Dynamics of Water Fluxes and Storages in an Alpine Karst Catchment under Current and Potential Future Climate Conditions". In: *Hydrol. Earth Syst. Sci.*, p. 17. DOI: 10.5194/hess-22-3807-2018.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. arXiv: 1409.1259 [cs, stat]. URL: http://arxiv.org/abs/1409.1259 (visited on 02/06/2022).

Chollet, F. (2015). *Keras*. URL: https://github.com/keras-team/keras (visited on 05/22/2020).

Cloutier, Claude-André, Thomas Buffin-Bélanger, and Marie Larocque (2014). "Controls of Groundwater Floodwave Propagation in a Gravelly Floodplain". In: *Journal of Hydrology* 511, pp. 423–431. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2014.02.014.

Collaud Coen, Martine, Elisabeth Andrews, Alessandro Bigi, Giovanni Martucci, Gonzague Romanens, Frédéric P. A. Vogt, and Laurent Vuilleumier (2020). "Effects of the Prewhitening Method, the Time Granularity, and the Time Segmentation on the Mann–Kendall Trend Detection and the Associated Sen's Slope". In: *Atmos. Meas. Tech.* 13.12, pp. 6945–6964. ISSN: 1867-8548. DOI: 10/gm2xc9.

CORINE Land Cover (2018). *CORINE Land Cover (CLC) — Copernicus Land Monitoring Service*. URL: https://land.copernicus.eu/user-corner/publications/clc-flyer (visited on 08/23/2019).

Cornes, Richard C., Gerard van der Schrier, Else J. M. van den Besselaar, and Philip D. Jones (2018). "An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets". In: *J. Geophys. Res. Atmos.* 123.17, pp. 9391–9409. ISSN: 2169897X. DOI: 10/gfk3hd.

Corona, Claudia R., Jason J. Gurdak, Jesse E. Dickinson, T.P.A. Ferré, and Edwin P. Maurer (2018). "Climate Variability and Vadose Zone Controls on Damping of Transient Recharge". In: *Journal of Hydrology* 561, pp. 1094–1104. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2017.08.028.

Cuthbert, M. O. (2014). "Straight Thinking about Groundwater Recession". In: *Water Resources Research* 50.3, pp. 2407–2424. ISSN: 1944-7973. DOI: 10.1002/2013wr014060.

Darras, T., V. Borrell Estupina, L. Kong-A-Siou, B. Vayssade, A. Johannet, and S. Pistre (2015). "Identification of Spatial and Temporal Contributions of Rainfalls to Flash Floods Using Neural Network Modelling: Case Study on the Lez Basin (Southern France)". In: *Hydrology and Earth System Sciences* 19.10, pp. 4397–4410. ISSN: 1027-5606. DOI: 10/f7xq6q.

Darras, Thomas, Line Kong-A-Siou, Bernard Vayssade, Anne Johannet, and Séverin Pistre (2017). "Karst Flash Flood Forecasting Using Recurrent and Nonrecurrent Artificial Neural Network Models: The Case of the Lez Basin (Southern France)". In: *EuroKarst 2016, Neuchâtel*. Advances in Karst Science. Cham: Springer International Publishing. ISBN: 978-3-319-45464-1 978-3-319-45465-8. DOI: 10.1007/978-3-319-45465-8.

De Graaf, Inge E. M., Tom Gleeson, L. P. H. (Rens) van Beek, Edwin H. Sutanudjaja, and Marc F. P. Bierkens (2019). "Environmental Flow Limits to Global Groundwater Pumping". In: *Nature* 574.7776, pp. 90–94. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1594-4.

Destatis (2021). *Wassergewinnung: Bundesländer, Jahre, Wasserarten*. GENESIS-Online. URL: https://www-genesis.destatis.de/genesis//online?operation=table&code=32211-0002&bypass=true&levelindex=0&levelid=1611589342283#abreadcrumb (visited on 01/25/2021).

Di Nunno, Fabio and Francesco Granata (2020). "Groundwater Level Prediction in Apulia Region (Southern Italy) Using NARX Neural Network". In: *Environmental Research* 190, p. 110062. ISSN: 0013-9351. DOI: 10.1016/j.envres.2020.110062.

Duan, Shiheng, Paul Ullrich, and Lele Shu (2020). "Using Convolutional Neural Networks for Streamflow Projection in California". In: *Front. Water* 2, p. 28. ISSN: 2624-9375. DOI: 10.3389/frwa.2020.00028.

DWD (2018). *Kern-Ensemble v2018*. URL: https://www.dwd.de/DE/klimaumwelt/klimaforschung/klimaprojektionen/fuer_deutschland/fuer_dtld_rcp-datensatz_node.html (visited on 02/17/2021).

DWD Climate Data Center (CDC) (2020). *Historical and Current Hourly RADOLAN Grids of Precipitation Depth (Binary). Version V001*. URL: https://opendata.dwd.de/climate_environment/CDC/grids_germany/hourly/radolan/ (visited on 12/11/2020).

Eckhardt, K. and U. Ulbrich (2003). "Potential Impacts of Climate Change on Groundwater Recharge and Streamflow in a Central European Low Mountain Range". In: *Journal of Hydrology* 284.1, pp. 244–252. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2003.08.005.

EEA (2017). *EU-Hydro — Copernicus Land Monitoring Service*. URL: https://land.copernicus.eu/user-corner/publications/eu-hydro-flyer (visited on 08/23/2019).

EURO-CORDEX (2018). *EURO-CORDEX Simulations*. URL: https://www.euro-cordex.net/060376/index.php.en (visited on 04/08/2021).

European Union (2022). *Copernicus*. Copernicus - Europe's eyes on Earth. URL: https://www.copernicus.eu/en (visited on 02/06/2022).

Fang, Kuai, Daniel Kifer, Kathryn Lawson, and Chaopeng Shen (2020). "Evaluating the Potential and Challenges of an Uncertainty Quantification Method for Long Short-Term Memory Models for Soil Moisture Predictions". In: *Water Resour. Res.* 56.12. ISSN: 0043-1397, 1944-7973. DOI: 10.1029/2020wr028095.

Fang, Kuai, Ming Pan, and Chaopeng Shen (2019). "The Value of SMAP for Long-Term Soil Moisture Estimation With the Help of Deep Learning". In: *IEEE Trans. Geosci. Remote Sensing* 57.4, pp. 2221–2233. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/tgrs.2018.2872131.

FAO (2010). *The Wealth of Waste: The Economics of Wastewater Use in Agriculture*. In collab. with J. T. Winpenny, I. Heinz, and S. Koo-Oshima. FAO Water Reports 35. Rome: Food and Agriculture Organization of the United Nations. 129 pp. ISBN: 978-92-5-106578-5.

Fleury, P., B. Ladouche, Y. Conroux, H. Jourde, and N. Dörfliger (2009). "Modelling the Hydrologic Functions of a Karst Aquifer under Active Water Management – The Lez Spring". In: *Journal of Hydrology* 365.3-4, pp. 235–243. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2008.11.037.

Fresnay, S., A. Hally, C. Garnaud, E. Richard, and D. Lambert (2012). "Heavy Precipitation Events in the Mediterranean: Sensitivity to Cloud Physics Parameterisation Uncertainties". In: *Nat. Hazards Earth Syst. Sci.* 12.8, pp. 2671–2688. ISSN: 1684-9981. DOI: 10.5194/nhess-12-2671-2012.

Frick, Claudia, Heiko Steiner, Alex Mazurkiewicz, Ulf Riediger, Monika Rauthe, Thomas Reich, and Annegret Gratzki (2014). "Central European High-Resolution Gridded Daily Data Sets (HYRAS): Mean Temperature and Relative Humidity". In: *Meteorologische Zeitschrift* 23.1, pp. 15–32. ISSN: 0941-2948. DOI: 10.1127/0941-2948/2014/0560.

Friedman, Jerome H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." In: *The Annals of Statistics* 29.5, pp. 1189–1232. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1013203451.

Fukushima, Kunihiko and Sei Miyake (1982). "Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position". In: *Pattern Recognition* 15.6, pp. 455–469. ISSN: 00313203. DOI: 10.1016/0031-3203(82)90024-3.

Gao, Chao, Marco Gemmer, Xiaofan Zeng, Bo Liu, Buda Su, and Yuhua Wen (2010). "Projected Streamflow in the Huaihe River Basin (2010–2100) Using Artificial Neural Network". In: *Stoch Environ Res Risk Assess* 24.5, pp. 685–697. ISSN: 1436-3259. DOI: 10.1007/s00477-009-0355-6.

Gauch, Martin, Frederik Kratzert, Daniel Klotz, Grey Nearing, Jimmy Lin, and Sepp Hochreiter (2020). *Rainfall–Runoff Prediction at Multiple Timescales with a SingleLong Short-Term Memory Network*. preprint. Catchment hydrology/Modelling approaches. DOI: 10.5194/hess-2020-540.

Gauch, Martin, Juliane Mai, and Jimmy Lin (2021). "The Proper Care and Feeding of CAMELS: How Limited Training Data Affects Streamflow Prediction". In: *Environmental Modelling & Software* 135, p. 104926. ISSN: 13648152. DOI: 10.1016/j.envsoft.2020.104926. arXiv: 1911.07249.

Gerlings, Julie, Millie Søndergaard Jensen, and Arisa Shollo (2022). "Explainable AI, But Explainable to Whom? An Exploratory Case Study of xAI in Healthcare". In: *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects*. Ed. by Chee-Peng Lim, Yen-Wei Chen, Ashlesha Vaidya, Charu Mahorkar, and Lakhmi C. Jain. Intelligent Systems Reference Library. Cham:

Springer International Publishing, pp. 169–198. ISBN: 978-3-030-83620-7. DOI: 10.1007/978-3-030-83620-7_7.

Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins (2000). "Learning to Forget: Continual Prediction with LSTM". In: *Neural Computation* 12.10, pp. 2451–2471. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976600300015015.

Geyer, Otto Franz, Manfred P. Gwinner, Matthias Geyer, Edgar Nitsch, Theo Simon, and Dietrich Ellwanger (2011). *Geologie von Baden-Württemberg*. 5., völlig neu bearb. Aufl. Stuttgart: Schweizerbart. 627 pp. ISBN: 978-3-510-65267-9.

Ghazi, Babak, Esmaeil Jeihouni, and Zahra Kalantari (2021). "Predicting Groundwater Level Fluctuations under Climate Change Scenarios for Tasuj Plain, Iran". In: *Arab J Geosci* 14.2, p. 115. ISSN: 1866-7538. DOI: 10.1007/s12517-021-06508-6.

Gholami, V., M. R. Khaleghi, S. Pirasteh, and Martijn J. Booij (2021). "Comparison of Self-Organizing Map, Artificial Neural Network, and Co-Active Neuro-Fuzzy Inference System Methods in Simulating Groundwater Quality: Geospatial Artificial Intelligence". In: *Water Resour Manage*. ISSN: 1573-1650. DOI: 10.1007/s11269-021-02969-2.

Giese, Markus, Ezra Haaf, Benedikt Heudorfer, and Roland Barthel (2020). "Comparative Hydrogeology – Reference Analysis of Groundwater Dynamics from Neighbouring Observation Wells". In: *Hydrological Sciences Journal* 0.0, pp. 1–22. ISSN: 0262-6667. DOI: 10.1080/02626667.2020.1762888.

Goldscheider, Nico (2005). "Fold Structure and Underground Drainage Pattern in the Alpine Karst System Hochifen-Gottesacker". In: *Eclogae geol. Helv.* 98.1, pp. 1–17. ISSN: 0012-9402, 1420-9128. DOI: 10.1007/s00015-005-1143-z.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: The MIT Press. 775 pp. ISBN: 978-0-262-03561-3.

Gupta, Hoshin V., Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez (2009). "Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling". In: *Journal of Hydrology* 377.1, pp. 80–91. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2009.08.003.

Guzman, Sandra M., Joel O. Paz, and Mary Love M. Tagert (2017). "The Use of NARX Neural Networks to Forecast Daily Groundwater Levels". In: *Water Resources Management* 31.5, pp. 1591–1603. ISSN: 0920-4741, 1573-1650. DOI: 10.1007/s11269-017-1598-5.

Guzman, Sandra M., Joel O. Paz, Mary Love M. Tagert, and Andrew E. Mercer (2019). "Evaluation of Seasonally Classified Inputs for the Prediction of Daily Groundwater Levels: NARX Networks Vs Support Vector Machines". In: *Environ. Model. Assess.* 24.2, pp. 223–234. ISSN: 1420-2026. DOI: 10.1007/s10666-018-9639-x.

Haaf, Ezra and Roland Barthel (2018). "An Inter-Comparison of Similarity-Based Methods for Organisation and Classification of Groundwater Hydrographs". In: *Journal of Hydrology* 559, pp. 222–237. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2018.02.035.

Hagan, Martin T., Howard B. Demuth, Mark Hudson Beale, and Orlando De Jésus (2014). *Neural Network Design*. 2nd edition. s.L: Martin T. Hagan. 1 p. ISBN: 978-0-9717321-1-7.

Han, Jing-Cheng, Yuefei Huang, Zhong Li, Chunhong Zhao, Guanhui Cheng, and Pengfei Huang (2016). "Groundwater Level Prediction Using a SOM-aided Stepwise Cluster Inference Model". In: *Journal of Environmental Management* 182, pp. 308–321. ISSN: 0301-4797. DOI: 10.1016/j.jenvman.2016.07.069.

Hasda, Ripon, Md. Ferozur Rahaman, Chowdhury Sarwar Jahan, Khademul Islam Molla, and Quamrul Hasan Mazumder (2020). "Climatic Data Analysis for Groundwater Level Simulation in Drought Prone Barind Tract, Bangladesh: Modelling Approach Using Artificial Neural Network". In: *Groundwater for Sustainable Development* 10, p. 100361. ISSN: 2352-801X. DOI: 10.1016/j.gsd.2020.100361.

He, Ji, Ah-Hwee Tan, Chew-Lim Tan, and Sam-Yuan Sung (2004). "On Quantitative Evaluation of Clustering Systems". In: Wu, Weili, Hui Xiong, and Shashi Shekhar. *Clustering and Information*

Retrieval. Red. by Ding-Zhu Du and Cauligi Raghavendra. Vol. 11. Network Theory and Applications. Boston, MA: Springer US, pp. 105–133. ISBN: 978-1-4613-7949-2 978-1-4613-0227-8. DOI: 10.1007/978-1-4613-0227-8_4.

Hebb, D.O (1949). "Organization of Behavior". In: *Journal of Clinical Psychology* 6.3, pp. 307–307. ISSN: 00219762, 10974679. DOI: 10.1002/1097-4679(195007)6:3<307::aid-jclp2270060338>3.0.co;2-k.

Herrmann, Frank, Ralf Kunkel, Ulrich Ostermann, Harry Vereecken, and Frank Wendland (2016). "Projected Impact of Climate Change on Irrigation Needs and Groundwater Resources in the Metropolitan Area of Hamburg (Germany)". In: *Environ Earth Sci* 75.14, p. 1104. ISSN: 1866-6299. DOI: 10.1007/s12665-016-5904-y.

Heudorfer, B., E. Haaf, K. Stahl, and R. Barthel (2019). "Index-based Characterization and Quantification of Groundwater Dynamics". In: *Water Resour. Res.* 55.7, pp. 5575–5592. ISSN: 0043-1397, 1944-7973. DOI: 10.1029/2018wr024418.

HLNUG (2019). *GruSchu*. Fachinformationssystem Grund- und Trinkwasserschutz Hessen. URL: http://gruschu.hessen.de (visited on 01/10/2018).

Hochreiter, S (1991). "Untersuchungen Zu Dynamischen Neuronalen Netzen". Munich: TU Munich. URL: http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1997.9.8.1735.

Hock, Regine (1999). "A Distributed Temperature-Index Ice- and Snowmelt Model Including Potential Direct Solar Radiation". In: *Journal of Glaciology* 45.149, pp. 101–111. ISSN: 0022-1430, 1727-5652. DOI: 10/ggnvkt.

Hölting, Bernward and Wilhelm Georg Coldewey (2013). *Hydrogeologie: Einführung in die allgemeine und angewandte Hydrogeologie*. 8. Aufl. Berlin: Springer-Spektrum. 438 pp. ISBN: 978-3-8274-2353-5 978-3-8274-2354-2.

Huebener, Heike, Katharina Bülow, Cornelia Fooken, Barbara Früh, Peter Hoffmann, Simona Höpp, Klaus Keuler, Christoph Menz, Viktoria Mohr, Kai Radtke, Hans Ramthun, Arne Spekat, Christian Steger, Frank Toussaint, Kirsten Warrach-Sagi, and Michael Woldt (2017). "ReKliEs-De Ergebnisbericht". In: DOI: 10.2312/WDCC/REKLIESDE_ERGEBNISBERICHT.

Hunter, John D. (2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3, pp. 90–95. ISSN: 1558-366X. DOI: 10.1109/mcse.2007.55.

Hussain, Dostdar, Tahir Hussain, Aftab Ahmed Khan, Syed Ali Asad Naqvi, and Akhtar Jamil (2020). "A Deep Learning Approach for Hydrological Time-Series Prediction: A Case Study of Gilgit River Basin". In: *Earth Sci Inform* 13.3, pp. 915–927. ISSN: 1865-0481. DOI: 10.1007/s12145-020-00477-2.

Hussain, Md. and Ishtiak Mahmud (2019). "pyMannKendall: A Python Package for Non Parametric Mann Kendall Family of Trend Tests." In: *JOSS* 4.39, p. 1556. ISSN: 2475-9066. DOI: 10/gjp7sn.

Hyndman, Rob J. and . Athanasopoulos (2021). *Forecasting: Principles and Practice (3rd Ed)*. OTexts: Melbourne. URL: https://0texts.com/fpp3/ (visited on 02/22/2022).

Idrizovic, Dzenita, Vesna Pocuca, Mirjam Vujadinovic Mandic, Nevenka Djurovic, Gordana Matovic, and Enika Gregoric (2020). "Impact of Climate Change on Water Resource Availability in a Mountainous Catchment: A Case Study of the Toplica River Catchment, Serbia". In: *Journal of Hydrology* 587, p. 124992. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2020.124992.

Innamorati, Carlo, Tobias Ritschel, Tim Weyrich, and Niloy J. Mitra (2020). "Learning on the Edge: Investigating Boundary Filters in CNNs". In: *Int J Comput Vis* 128.4, pp. 773–782. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01223-y.

Ioffe, Sergey and Christian Szegedy (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. arXiv: 1502.03167 [cs]. URL: http://arxiv.org/abs/1502.03167 (visited on 11/14/2021).

IPCC (2014). *Climate Change 2014: Synthesis Report*. Ed. by R. K. Pachauri and Leo Mayer. Geneva, Switzerland: Intergovernmental Panel on Climate Change. 151 pp. ISBN: 978-92-9169-143-2.

Izady, A., K. Davary, A. Alizadeh, A. Moghaddamnia, A. N. Ziaei, and S. M. Hasheminia (2013). "Application of NN-ARX Model to Predict Groundwater Levels in the Neishaboor Plain, Iran". In: *Water Resources Management* 27.14, pp. 4773–4794. ISSN: 0920-4741, 1573-1650. DOI: 10.1007/s11269-013-0432-y.

Jacob, Daniela, Juliane Petersen, Bastian Eggert, Antoinette Alias, Ole Bø ssing Christensen, Laurens M. Bouwer, Alain Braun, Augustin Colette, Michel Déqué, Goran Georgievski, Elena Georgopoulou, Andreas Gobiet, Laurent Menut, Grigory Nikulin, Andreas Haensler, Nils Hempelmann, Colin Jones, Klaus Keuler, Sari Kovats, Nico Kröner, Sven Kotlarski, Arne Kriegsmann, Eric Martin, Erik van Meijgaard, Christopher Moseley, Susanne Pfeifer, Swantje Preuschmann, Christine Radermacher, Kai Radtke, Diana Rechid, Mark Rounsevell, Patrick Samuelsson, Samuel Somot, Jean-Francois Soussana, Claas Teichmann, Riccardo Valentini, Robert Vautard, Björn Weber, and Pascal Yiou (2014). "EURO-CORDEX: New High-Resolution Climate Change Projections for European Impact Research". In: *Reg Environ Change* 14.2, pp. 563–578. ISSN: 1436-378X. DOI: 10/f9sfkm.

Jakeman, A. J. and G. M. Hornberger (1993). "How Much Complexity Is Warranted in a Rainfall-Runoff Model?" In: *Water Resources Research* 29.8, pp. 2637–2649. ISSN: 1944-7973. DOI: 10.1029/93wr00877.

Jasechko, Scott, S. Jean Birks, Tom Gleeson, Yoshihide Wada, Peter J. Fawcett, Zachary D. Sharp, Jeffrey J. McDonnell, and Jeffrey M. Welker (2014). "The Pronounced Seasonality of Global Groundwater Recharge". In: *Water Resour. Res.* 50.11, pp. 8845–8867. ISSN: 00431397. DOI: 10.1002/2014WR015809.

Jeannin, Pierre-Yves, Guillaume Artigue, Christoph Butscher, Yong Chang, Jean-Baptiste Charlier, Lea Duran, Laurence Gill, Andreas Hartmann, Anne Johannet, Hervé Jourde, Alireza Kavousi, Tanja Liesch, Yan Liu, Martin Lüthi, Arnauld Malard, Naomi Mazzilli, Eulogio Pardo-Igúzquiza, Dominique Thiéry, Thomas Reimann, Philip Schuler, Thomas Wöhling, and Andreas Wunsch (2021). "Karst Modelling Challenge 1: Results of Hydrological Modelling". In: *Journal of Hydrology* 600, p. 126508. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2021.126508.

Jeihouni, Esmaeil, Saeid Eslamian, Mirali Mohammadi, and Mohammad Javad Zareian (2019a). "Simulation of Groundwater Level Fluctuations in Response to Main Climate Parameters Using a Wavelet–ANN Hybrid Technique for the Shabestar Plain, Iran". In: *Environ Earth Sci* 78.10, p. 293. ISSN: 1866-6280, 1866-6299. DOI: 10.1007/s12665-019-8283-3.

Jeihouni, Esmaeil, Mirali Mohammadi, Saeid Eslamian, and Mohammad Javad Zareian (2019b). "Potential Impacts of Climate Change on Groundwater Level through Hybrid Soft-Computing Methods: A Case Study—Shabestar Plain, Iran". In: *Environ Monit Assess* 191.10, p. 620. ISSN: 1573-2959. DOI: 10.1007/s10661-019-7784-6.

Jeong, Jina and Eungyu Park (2019). "Comparative Applications of Data-Driven Models Representing Water Table Fluctuations". In: *Journal of Hydrology* 572, pp. 261–273. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2019.02.051.

Jeong, Jina, Eungyu Park, Huali Chen, Kue-Young Kim, Weon Shik Han, and Heejun Suk (2020). "Estimation of Groundwater Level Based on the Robust Training of Recurrent Neural Networks Using Corrupted Data". In: *Journal of Hydrology* 582, p. 124512. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2019.124512.

Johannet, Anne, Alain Mangin, and D. D'Hulst (1994). "Subterranean Water Infiltration Modelling by Neural Networks: Use of Water Source Flow". In: *ICANN '94: Proceedings of the International Conference on Artificial Neural Networks Sorrento, Italy, 26–29 May 1994 Volume 1, Parts 1 and 2*. International Conference on Artificial Neural Networks. Sorrento, Italy: Springer Berlin Heidelberg, pp. 1033–1036. ISBN: 978-3-540-19887-1.

Jourde, H., N. Massei, N. Mazzilli, S. Binet, C. Batiot-Guilhe, D. Labat, M. Steinmann, V. Bailly-Comte, J. L. Seidel, B. Arfib, J. B. Charlier, V. Guinot, A. Jardani, M. Fournier, M. Aliouache, M. Babic, C. Bertrand, P. Brunet, J. F. Boyer, J. P. Bricquet, T. Camboulive, S. D. Carrière, H. Celle-Jeanton, K. Chalikakis, N. Chen, C. Cholet, V. Clauzon, L. Dal Soglio, C. Danquigny, C. Défargue, S. Denimal, C. Emblanch, F. Hernandez, M. Gillon, A. Gutierrez, L. Hidalgo Sanchez, M. Hery, N. Houillon, A. Johannet, J. Jouves, N. Jozja, B. Ladouche, V. Leonardi, G. Lorette,

C. Loup, P. Marchand, V. de Montety, R. Muller, C. Ollivier, V. Sivelle, R. Lastennet, N. Lecoq, J. C. Maréchal, L. Perotin, J. Perrin, M. A. Petre, N. Peyraube, S. Pistre, V. Plagnes, A. Probst, J. L. Probst, R. Simler, V. Stefani, D. Valdes-Lao, S. Viseur, and X. Wang (2018). "SNO KARST: A French Network of Observatories for the Multidisciplinary Study of Critical Zone Processes in Karst Watersheds and Aquifers". In: *Vadose Zone Journal* 17.1, p. 180094. ISSN: 1539-1663. DOI: 10/gk9t3n.

Jourde, Hervé, A. Lafare, N. Mazzilli, G. Belaud, L. Neppel, N. Dörfliger, and F. Cernesson (2014). "Flash Flood Mitigation as a Positive Consequence of Anthropogenic Forcing on the Groundwater Resource in a Karst Catchment". In: *Environ Earth Sci* 71.2, pp. 573–583. ISSN: 1866-6280, 1866-6299. DOI: 10.1007/s12665-013-2678-3.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis (2021). "Highly Accurate Protein Structure Prediction with AlphaFold". In: *Nature* 596.7873 (7873), pp. 583–589. ISSN: 1476-4687. DOI: 10/gk7nfp.

Karniadakis, George Em, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang (2021). "Physics-Informed Machine Learning". In: *Nat Rev Phys* 3.6 (6), pp. 422–440. ISSN: 2522-5820. DOI: 10.1038/s42254-021-00314-5.

Kaufman, Georg, Cyril Mayaud, Blaž Kogovšek, and Franci Gabrovšek (2020). "Understanding the Temporal Variation of Flow Direction in a Complex Karst System (Planinska Jama, Slovenia)". In: *Acta Carsologica* 49.2-3 (2-3). ISSN: 1580-2612. DOI: 10/gkcs9h.

Kaufmann, Georg, Franci Gabrovšek, and Janez Turk (2016). "Modelling Flow of Subterranean Pivka River in Postojnska Jama, Slovenia". In: *Acta Carsologica* 45.1 (1). ISSN: 1580-2612. DOI: 10.3986/ac.v45i1.3059.

Kersebaum, K. C. and C. Nendel (2014). "Site-Specific Impacts of Climate Change on Wheat Production across Regions of Germany Using Different CO2 Response Functions". In: *European Journal of Agronomy*. Land, Climate and Resources 2020. Decision Support for Agriculture under Climate Change 52, pp. 22–32. ISSN: 1161-0301. DOI: 10.1016/j.eja.2013.04.005.

Kiang, Melody Y., Michael Y. Hu, and Dorothy M. Fisher (2006). "An Extended Self-Organizing Map Network for Market Segmentation—a Telecommunication Example". In: *Decision Support Systems* 42.1, pp. 36–47. ISSN: 01679236. DOI: 10.1016/j.dss.2004.09.012.

King, David A., Dominique M. Bachelet, Amy J. Symstad, Ken Ferschweiler, and Michael Hobbins (2015). "Estimation of Potential Evapotranspiration from Extraterrestrial Radiation, Air Temperature and Humidity to Assess Future Climate Change Effects on the Vegetation of the Northern Great Plains, USA". In: *Ecological Modelling* 297, pp. 86–97. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2014.10.037.

Kiranyaz, Serkan, Turker Ince, Osama Abdeljaber, Onur Avci, and Moncef Gabbouj (2019). "1-D Convolutional Neural Networks for Signal Processing Applications". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8360–8364. DOI: 10.1109/ICASSP.2019.8682194.

Klotz, Daniel, Frederik Kratzert, Martin Gauch, Alden Keefe Sampson, Günter Klambauer, Sepp Hochreiter, and Grey Nearing (2020). *Uncertainty Estimation with Deep Learning for Rainfall-Runoff Modelling*. arXiv: 2012.14295 [physics]. URL: http://arxiv.org/abs/2012.14295 (visited on 03/30/2021).

Kløve, Bjørn, Pertti Ala-Aho, Guillaume Bertrand, Jason J. Gurdak, Hans Kupfersberger, Jens Kværner, Timo Muotka, Heikki Mykrä, Elena Preda, Pekka Rossi, Cintia Bertacchi Uvo, Elzie Velasco, and Manuel Pulido-Velazquez (2014). "Climate Change Impacts on Groundwater and Dependent

Ecosystems". In: *Journal of Hydrology*. Climatic Change Impact on Water: Overcoming Data and Science Gaps 518, pp. 250–266. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2013.06.037.

Kohavi, Ron and George H. John (1997). "Wrappers for Feature Subset Selection". In: *Artificial Intelligence* 97.1-2, pp. 273–324. ISSN: 00043702. DOI: 10.1016/S0004-3702(97)00043-X.

Kohonen, Teuvo (2014). *Matlab Implementations and Applications of the Self-Organizing Map*. Helsinki. ISBN: 978-952-60-3679-3.

Kollat, J. B., P. M. Reed, and T. Wagener (2012). "When Are Multiobjective Calibration Trade-Offs in Hydrologic Models Meaningful?" In: *Water Resources Research* 48.3. ISSN: 1944-7973. DOI: 10/gkcq5k.

Kong-A-Siou, Line, Kévin Cros, Anne Johannet, Valérie Borrell-Estupina, and Séverin Pistre (2013). "KnoX Method, or Knowledge eXtraction from Neural Network Model. Case Study on the Lez Karst Aquifer (Southern France)". In: *Journal of Hydrology* 507, pp. 19–32. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2013.10.011.

Kong-A-Siou, Line, Perrine Fleury, Anne Johannet, Valérie Borrell Estupina, Séverin Pistre, and Nathalie Dörfliger (2014). "Performance and Complementarity of Two Systemic Models (Reservoir and Neural Networks) Used to Simulate Spring Discharge and Piezometry for a Karst Aquifer". In: *Journal of Hydrology* 519, pp. 3178–3192. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2014.10.041.

Kong A Siou, Line, Anne Johannet, Valérie Borrell, and Séverin Pistre (2011). "Complexity Selection of a Neural Network Model for Karst Flood Forecasting: The Case of the Lez Basin (Southern France)". In: *Journal of Hydrology* 403.3-4, pp. 367–380. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2011.04.015.

Kong-A-Siou, Line, Anne Johannet, Valérie Borrell Estupina, and Séverin Pistre (2015). "Neural Networks for Karst Groundwater Management: Case of the Lez Spring (Southern France)". In: *Environmental Earth Sciences* 74.12, pp. 7617–7632. ISSN: 1866-6280, 1866-6299. DOI: 10.1007/s12665-015-4708-9.

Kong A Siou, Line, Anne Johannet, Borrell Estupina Valérie, and Séverin Pistre (2012). "Optimization of the Generalization Capability for Rainfall–Runoff Modeling by Neural Networks: The Case of the Lez Aquifer (Southern France)". In: *Environmental Earth Sciences* 65.8, pp. 2365–2375. ISSN: 1866-6280, 1866-6299. DOI: 10.1007/s12665-011-1450-9.

Kovačič, Gregor, Metka Petrič, and Nataša Ravbar (2020). "Evaluation and Quantification of the Effects of Climate and Vegetation Cover Change on Karst Water Sources: Case Studies of Two Springs in South-Western Slovenia". In: *Water* 12.11 (11), p. 3087. DOI: 10/gksnmz.

Kraft, B., M. Jung, M. Körner, and M. Reichstein (2020). "Hybrid Modeling: Fusion of a Deep Learning Approach and a Physics-Based Model for Global Hydrological Modeling". In: *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLIII-B2-2020, pp. 1537–1544. ISSN: 2194-9034. DOI: 10.5194/isprs-archives-xliii-b2-2020-1537-2020.

Kratzert, Frederik, Mathew Herrnegger, Daniel Klotz, Sepp Hochreiter, and Günter Klambauer (2019a). *NeuralHydrology - Interpreting LSTMs in Hydrology*. arXiv: 1903.07903 [physics, stat]. URL: http://arxiv.org/abs/1903.07903 (visited on 10/31/2019).

Kratzert, Frederik, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger (2018). "Rainfall–Runoff Modelling Using Long Short-Term Memory (LSTM) Networks". In: *Hydrology and Earth System Sciences* 22.11, pp. 6005–6022. ISSN: 1027-5606. DOI: 10.5194/hess-22-6005-2018.

Kratzert, Frederik, Daniel Klotz, Guy Shalev, Günter Klambauer, Sepp Hochreiter, and Grey Nearing (2019b). "Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine Learning Applied to Large-Sample Datasets". In: *Hydrol. Earth Syst. Sci.* 23.12, pp. 5089–5110. ISSN: 1607-7938. DOI: 10.5194/hess-23-5089-2019.

Kreienkamp, Frank, Heike Huebener, Carsten Linke, and Arne Spekat (2012). "Good Practice for the Usage of Climate Model Simulation Results - a Discussion Paper". In: *Environ Syst Res* 1.1, p. 9. ISSN: 2193-2697. DOI: 10.1186/2193-2697-1-9.

Kreins, Peter, Martin Henseler, Jano Anter, Frank Herrmann, and Frank Wendland (2015). "Quantification of Climate Change Impact on Regional Agricultural Irrigation and Groundwater Demand". In: *Water Resour Manage* 29.10, pp. 3585–3600. ISSN: 1573-1650. DOI: 10.1007/s11269-015-1017-8.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc. URL: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 02/03/2022).

Kumar, Usha A. and Yuvnish Dhamija (2010). "Comparative Analysis of SOM Neural Network with K-means Clustering Algorithm". In: *2010 IEEE International Conference on Management of Innovation Technology*. 2010 IEEE International Conference on Management of Innovation Technology, pp. 55–59. DOI: 10.1109/icmit.2010.5492838.

Lähivaara, Timo, Alireza Malehmir, Antti Pasanen, Leo Kärkkäinen, Janne M. J. Huttunen, and Jan S. Hesthaven (2019). "Estimation of Groundwater Storage from Seismic Data Using Deep Learning". In: *Geophysical Prospecting* 67.8, pp. 2115–2126. ISSN: 1365-2478. DOI: 10.1111/1365-2478.12831.

Lam, A., D. Karssenberg, B. J. J. M. van den Hurk, and M. F. P. Bierkens (2011). "Spatial and Temporal Connections in Groundwater Contribution to Evaporation". In: *Hydrol. Earth Syst. Sci.* 15.8, pp. 2621–2630. ISSN: 1607-7938. DOI: 10.5194/hess-15-2621-2011.

Lebigot, Eric O. (2010–2020). *Uncertainties: A Python Package for Calculations with Uncertainties*. Version 3.0.1. URL: https://pythonhosted.org/uncertainties/numpy_guide.html (visited on 02/18/2021).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep Learning". In: *Nature* 521.7553, pp. 436–444. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539.

Lee, Daeeop, Giha Lee, Seongwon Kim, and Sungho Jung (2020). "Future Runoff Analysis in the Mekong River Basin under a Climate Change Scenario Using Deep Learning". In: *Water* 12.6, p. 1556. ISSN: 2073-4441. DOI: 10.3390/w12061556.

LGRB (2007). *Hydrogeologischer Bau und Aquifereigenschaften der Lockergesteine im Oberrheingraben (Baden-Württemberg)*. In collab. with Gunther Wirsing and Alexander Luz. URL: https://lgrb-bw.de/hydrogeologie/projekte/org.

Li, Qing, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen (2014). "Medical Image Classification with Convolutional Neural Network". In: *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*. 2014 13th International Conference on Control Automation Robotics Vision (ICARCV), pp. 844–848. DOI: 10.1109/ICARCV.2014.7064414.

Lin, Gwo-Fong and Lu-Hsien Chen (2005). "Time Series Forecasting by Combining the Radial Basis Function Network and the Self-Organizing Map". In: *Hydrological Processes* 19.10, pp. 1925–1937. ISSN: 1099-1085. DOI: 10.1002/hyp.5637.

Lin, Tsungnan, Bill G. Horne, and C. Lee Giles (1998). "How Embedded Memory in Recurrent Neural Network Architectures Helps Learning Long-Term Temporal Dependencies". In: *Neural Networks* 11.5, pp. 861–868. DOI: 10.1016/s0893-6080(98)00018-5.

Lin, Tsungnan, Bill G. Horne, Peter Tiňo, and C. Lee Giles (1996). "Learning Long-Term Dependencies in NARX Recurrent Neural Networks". In: *IEEE Transactions on Neural Networks* 7.6, pp. 1329–1338. DOI: 10.1109/72.548162.

Lin, Tsungnan, Bill G. Horne, Peter Tiňo, and C. Lee Giles (1995). "Learning Long-Term Dependencies Is Not as Difficult with NARX Networks". In: *Proceedings of the 8th International Conference on Neural Information Processing Systems* (Denver, Colorado). NIPS'95. Cambridge, MA, USA: MIT Press, pp. 577–583.

Longenecker, Jake, Timothy Bechtel, Zhao Chen, Nico Goldscheider, Tanja Liesch, and Robert Walter (2017). "Correlating Global Precipitation Measurement Satellite Data with Karst Spring Hydrographs for Rapid Catchment Delineation". In: *Geophysical Research Letters* 44.10, pp. 4926–4932. ISSN: 1944-8007. DOI: 10.1002/2017GL073790.

Longuevergne, Laurent, Nicolas Florsch, and Philippe Elsass (2007). "Extracting Coherent Regional Information from Local Measurements with Karhunen-Loève Transform: Case Study of an Alluvial Aquifer (Rhine Valley, France and Germany)". In: *Water Resources Research* 43.4. ISSN: 00431397. DOI: 10.1029/2006wr005000.

Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts (2013). "Understanding Variable Importances in Forests of Randomized Trees". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc. URL: https://papers.nips.cc/paper/2013/hash/e3796ae838835da0b6f6ea37bcf8bcb7-Abstract.html (visited on 01/30/2022).

LUBW (2006). *Hydrogeologischer Bau Und Hydraulische Eigenschaften - 9INTERREG III A-Projekt MoNit "Modellierung Der Grundwasserbelastung Durch Nitrat Im Oberrheingraben" / Structure Hydrogéologique et Caractéristiques Hydrauliques - 9INTERREG III A : MoNit "Modélisation de La Pollution Des Eaux Souterraines Par Les Nitrates Dans La Vallée Du Rhin Supérieur"*. LUBW. URL: https://pudi.lubw.de/detailseite/-/publication/37116-INTERREG_III_A-Projekt_MoNit__Modellierung_der_Grundwasserbelastung_durch_Nitrat_im_Oberrheingraben_.pdf (visited on 06/11/2019).

– (2018). *UDO - Umwelt-Daten und -Karten Online*. URL: https://udo.lubw.baden-wuerttemberg.de/public/ (visited on 01/12/2018).

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., pp. 4765–4774. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Machiwal, Deepesh and P. K. Singh (2015). "Understanding Factors Influencing Groundwater Levels in Hard-Rock Aquifer Systems by Using Multivariate Statistical Techniques". In: *Environ Earth Sci* 74.7, pp. 5639–5652. ISSN: 1866-6280, 1866-6299. DOI: 10.1007/s12665-015-4578-1.

Maier, Holger R. (1995). "Use of Artificial Neural Networks for Modelling Multivariate Water Quality Time Series". University of Adelaide.

Maier, Holger R. and Graeme C. Dandy (2000). "Neural Networks for the Prediction and Forecasting of Water Resources Variables: A Review of Modelling Issues and Applications". In: *Environmental modelling & software* 15.1, pp. 101–124. DOI: 10.1016/s1364-8152(99)00007-9.

Maier, Holger R., Ashu Jain, Graeme C. Dandy, and K.P. Sudheer (2010). "Methods Used for the Development of Neural Networks for the Prediction of Water Resource Variables in River Systems: Current Status and Future Directions". In: *Environmental Modelling & Software* 25.8, pp. 891–909. ISSN: 13648152. DOI: 10.1016/j.envsoft.2010.02.003.

Malard, Arnauld, Pierre-Yves Jeannin, Jonathan Vouillamoz, and Eric Weber (2015). "An Integrated Approach for Catchment Delineation and Conduit-Network Modeling in Karst Aquifers: Application to a Site in the Swiss Tabular Jura". In: *Hydrogeol J* 23.7, pp. 1341–1357. ISSN: 1435-0157. DOI: 10.1007/s10040-015-1287-5.

Mangiameli, Paul, Shaw K. Chen, and David West (1996). "A Comparison of SOM Neural Network and Hierarchical Clustering Methods". In: *European Journal of Operational Research* 93.2, pp. 402–417. ISSN: 03772217. DOI: 10.1016/0377-2217(96)00038-0.

Marx, Andreas, Rohini Kumar, Stephan Thober, Matthias Zink, Niko Wanders, Eric F. Wood, Ming Pan, Sheffield Justin, and Luis Samaniego (2017). "Climate Change Alters Low Flows in Europe under a 1.5, 2, and 3 Degree Global Warming". In: p. 24.

Mathworks Inc. (2020). *Matlab 2020a*. Version 2020a.

Mayaud, Cyril, Franci Gabrovšek, Matej Blatnik, Blaž Kogovšek, Metka Petrič, and Nataša Ravbar (2019). "Understanding Flooding in Poljes: A Modelling Perspective". In: *Journal of Hydrology* 575, pp. 874–889. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2019.04.092.

Mazzilli, Naomi, Hervé Jourde, Vincent Guinot, Vincent Bailly-Comte, and Perrine Fleury (2011). "Hydrological Modelling of a Karst Aquifer under Active Groundwater Management Using a Parsimonious Conceptual Model". In: *H2Karst*. Besançon, France. URL: https://hal.archives-ouvertes.fr/hal-01844603 (visited on 07/28/2021).

McCulloch, Warren S. and Walter Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/bf02478259.

McGovern, Amy, Ryan Lagerquist, David John Gagne, G. Eli Jergensen, Kimberly L. Elmore, Cameron R. Homeyer, and Travis Smith (2019). "Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning". In: *Bulletin of the American Meteorological Society* 100.11, pp. 2175–2199. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-D-18-0195.1.

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: Python in Science Conference. Austin, Texas, pp. 56–61. DOI: 10.25080/majora-92bf1922-00a.

McMillan, Hilary, Ida Westerberg, and Flora Branger (2017). "Five Guidelines for Selecting Hydrological Signatures". In: *Hydrological Processes* 31.26, pp. 4757–4761. ISSN: 1099-1085. DOI: 10.1002/hyp.11300.

Mekonnen, Mesfin M. and Arjen Y. Hoekstra (2016). "Four Billion People Facing Severe Water Scarcity". In: *Sci. Adv.* 2.2, e1500323. ISSN: 2375-2548. DOI: 10/gc6xww.

Melo Riveros, Nicolas Andres, Bayron Alexis Cardenas Espitia, and Lilia Edith Aparicio Pico (2019). "Comparison between K-means and Self-Organizing Maps Algorithms Used for Diagnosis Spinal Column Patients". In: *Informatics in Medicine Unlocked* 16, p. 100206. ISSN: 23529148. DOI: 10.1016/j.imu.2019.100206.

Mingoti, Sueli A. and Joab O. Lima (2006). "Comparing SOM Neural Network with Fuzzy C-Means, K-means and Traditional Hierarchical Clustering Algorithms". In: *European Journal of Operational Research* 174.3, pp. 1742–1759. ISSN: 03772217. DOI: 10.1016/j.ejor.2005.03.039.

Mitchell, Tom M. (1980). *The Need for Biases in Learning Generalizations*. New Brunswick, NJ: Rutgers University. URL: http://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf.

Moghaddamnia, A., R. Remesan, M. Hassanpour Kashani, M. Mohammadi, D. Han, and J. Piri (2009). "Comparison of LLR, MLP, Elman, NNARX and ANFIS Models—with a Case Study in Solar Radiation Estimation". In: *Journal of Atmospheric and Solar-Terrestrial Physics* 71.8-9, pp. 975–982. ISSN: 13646826. DOI: 10.1016/j.jastp.2009.04.009.

Moradkhani, Hamid, Kuo-lin Hsu, Hoshin V. Gupta, and Soroosh Sorooshian (2004). "Improved Streamflow Forecasting Using Self-Organizing Radial Basis Function Artificial Neural Networks". In: *Journal of Hydrology* 295.1, pp. 246–262. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2004.03.027.

Moss, Richard, Mustafa Babiker, Sander Brinkman, Eduardo Calvo, Tim Carter, Jae Edmonds, Ismail Elgizouli, Seita Emori, Lin Erda, Kathy Hibbard, Roger Jones, Mikiko Kainuma, Jessica Kelleher, Jean Francois Lamarque, Martin Manning, Ben Matthews, Jerry Meehl, Leo Meyer, John Mitchell, Nebojsa Nakicenovic, Brian O'Neill, Ramon Pichs, Keywan Riahi, Steven Rose, Paul Runci, Ron Stouffer, Detlef van Vuuren, John Weyant, Tom Wilbanks, Jean Pascal van van Ypersele, and Monika Zurek (2008). *Towards New Scenarios for Analysis of Emissions, Climate Change, Impacts, and Response Strategies*. Geneva: Intergovernmental Panel on Climate Change, p. 132. URL: http://ipcc-data.org/docs/ar5scenarios/IPCC_Final_Draft_Meeting_Report_3May08.pdf (visited on 04/08/2021).

Muñoz Sabater, J. (2019). *ERA5-Land Hourly Data from 2001 to Present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. DOI: 10.24381/CDS.E2161BAC.

MUEEF (2018). *Geoportal Wasser*. URL: http://geoportal-wasser.rlp.de/servlet/is/8183/ (visited on 07/07/2018).

Müller, Juliane, Jangho Park, Reetik Sahu, Charuleka Varadharajan, Bhavna Arora, Boris Faybishenko, and Deborah Agarwal (2020). "Surrogate Optimization of Deep Neural Networks for Groundwater Predictions". In: *J Glob Optim*. ISSN: 1573-2916. DOI: 10.1007/s10898-020-00912-0. arXiv: 1908.10947.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press. 1067 pp. ISBN: 978-0-262-01802-9.

Naranjo-Fernández, Nuria, Carolina Guardiola-Albert, Héctor Aguilera, Carmen Serrano-Hidalgo, and Esperanza Montero-González (2020). "Clustering Groundwater Level Time Series of the Exploited

Almonte-Marismas Aquifer in Southwest Spain". In: *Water* 12.4, p. 1063. ISSN: 2073-4441. DOI: 10.3390/w12041063.

NASA (2016). *GPM - Global Precipitation Measurement*. URL: http://www.nasa.gov/mission_pages/GPM/main/index.html (visited on 06/08/2021).

Nash, J. Eamonn and Jonh V. Sutcliffe (1970). "River Flow Forecasting through Conceptual Models Part I—A Discussion of Principles". In: *Journal of hydrology* 10.3, pp. 282–290. DOI: 10.1016/0022-1694(70)90255-6.

Neukum, Christoph and Rafig Azzam (2012). "Impact of Climate Change on Groundwater Recharge in a Small Catchment in the Black Forest, Germany". In: *Hydrogeol J* 20.3, pp. 547–560. ISSN: 1431-2174, 1435-0157. DOI: 10.1007/s10040-011-0827-x.

Nogueira, Fernando (2014). *Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python*. URL: https://github.com/fmfn/BayesianOptimization (visited on 04/15/2020).

Nourani, Vahid, Mohammad Taghi Alami, and Farnaz Daneshvar Vousoughi (2015). "Wavelet-Entropy Data Pre-Processing Approach for ANN-based Groundwater Level Modeling". In: *Journal of Hydrology* 524, pp. 255–269. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2015.02.048.

Pan, Mingyang, Hainan Zhou, Jiayi Cao, Yisai Liu, Jiangling Hao, Shaoxi Li, and Chi-Hua Chen (2020). "Water Level Prediction Model Based on GRU and CNN". In: *IEEE Access* 8, pp. 60090–60100. ISSN: 2169-3536. DOI: 10.1109/access.2020.2982433.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau (2011). "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Petrič, Metka, Janja Kogovšek, and Nataša Ravbar (2018). "Effects of the Vadose Zone on Groundwater Flow and Solute Transport Characteristics in Mountainous Karst Aquifers – the Case of the Javorniki–Snežnik Massif (SW Slovenia)". In: *AC* 47.1. ISSN: 0583-6050. DOI: 10.3986/ac.v47i1.5144.

Petsiuk, Vitali, Abir Das, and Kate Saenko (2018). *RISE: Randomized Input Sampling for Explanation of Black-box Models*. arXiv: 1806.07421 [cs]. URL: http://arxiv.org/abs/1806.07421 (visited on 11/13/2021).

Rahmani, Farshid, Kathryn Lawson, Wenyu Ouyang, Alison Appling, Samantha Oliver, and Chaopeng Shen (2021). "Exploring the Exceptional Performance of a Deep Learning Stream Temperature Model and the Value of Streamflow Data". In: *Environ. Res. Lett.* 16.2. ISSN: 1748-9326. DOI: 10.1088/1748-9326/abd501.

Rajaee, Taher, Hadi Ebrahimi, and Vahid Nourani (2019). "A Review of the Artificial Intelligence Methods in Groundwater Level Modeling". In: *Journal of Hydrology* 572, pp. 336–351. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2018.12.037.

Rauthe, Monika, Heiko Steiner, Ulf Riediger, Alex Mazurkiewicz, and Annegret Gratzki (2013). "A Central European Precipitation Climatology – Part I: Generation and Validation of a High-Resolution Gridded Daily Data Set (HYRAS)". In: *Meteorol. Z.*, p. 22. DOI: 10.1127/0941-2948/2013/0436.

Reback, Jeff, Wes McKinney, Jbrockmendel, Joris Van Den Bossche, Tom Augspurger, Phillip Cloud, Gfyoung, Sinhrks, Adam Klein, Matthew Roeschke, Simon Hawkins, Jeff Tratner, Chang She, William Ayd, Terji Petersen, Marc Garcia, Jeremy Schendel, Andy Hayden, MomIsBestFriend, Vytautas Jancauskas, Pietro Battiston, Skipper Seabold, Chris-B1, H-Vetinari, Stephan Hoyer, Wouter Overmeire, Alimcmaster1, Kaiqi Dong, Christopher Whelan, and Mortada Mehyar (2020). *Pandas-Dev/Pandas: Pandas 1.0.3*. Version v1.0.3. Zenodo. DOI: 10.5281/ZENODO.3509134.

Regierungspräsidium Darmstadt (1999). *Grundwasserbewirtschaftungsplan Hessisches Ried*.

Région Alsace - Strasbourg (1999). *Bestandsaufnahme Der Grundwasserqualität Im Oberrheingraben / Inventaire de La Qualité Des Eaux Souterraines Dans La Vallée Du Rhin Supérieur*. URL: https://www.ermes-rhin.eu/uploads/pdf/acces-libres/Resultats-INV1997.pdf (visited on 06/11/2019).

Richter, Brian D., Jeffrey V. Baumgartner, Jennifer Powell, and David P. Braun (1996). "A Method for Assessing Hydrologic Alteration within Ecosystems". In: *Conservation Biology* 10.4, pp. 1163–1174. ISSN: 0888-8892, 1523-1739. DOI: 10.1046/j.1523-1739.1996.10041163.x.

Rosenblatt, Frank (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." In: *Psychological review* 65.6, p. 386. DOI: 10.1037/h0042519.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning Representations by Back-Propagating Errors". In: *Nature* 323.6088, pp. 533–536. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/323533a0.

Schmidhuber, Juergen (2015). "Deep Learning in Neural Networks: An Overview". Version 4. In: *Neural Networks* 61, pp. 85–117. ISSN: 08936080. DOI: 10/f6v78n. arXiv: 1404.7828.

Schwalm, Christopher R., Spencer Glendon, and Philip B. Duffy (2020). "RCP8.5 Tracks Cumulative CO2 Emissions". In: *PNAS* 117.33, pp. 19656–19657. ISSN: 0027-8424, 1091-6490. DOI: 10/gg6xw2. pmid: 32747549.

Seibert, J. (2000). "Multi-Criteria Calibration of a Conceptual Runoff Model Using a Genetic Algorithm". In: *Hydrology and Earth System Sciences* 4.2, pp. 215–224. ISSN: 1027-5606. DOI: 10/crmzd7.

Sen, Pranab Kumar (1968). "Estimates of the Regression Coefficient Based on Kendall's Tau". In: *Journal of the American Statistical Association* 63.324, pp. 1379–1389. DOI: 10.1080/01621459.1968.10480934.

Sezen, Cenk, Nejc Bezak, Yun Bai, and Mojca Šraj (2019). "Hydrological Modelling of Karst Catchment Using Lumped Conceptual and Data Mining Models". In: *Journal of Hydrology* 576, pp. 98–110. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2019.06.036.

Shen, Chaopeng (2018a). "A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists". In: *Water Resour. Res.* 54.11, pp. 8558–8593. ISSN: 0043-1397, 1944-7973. DOI: 10.1029/2018wr022643.

– (2018b). "Deep Learning: A Next-Generation Big-Data Approach for Hydrology". In: *Eos* 99. ISSN: 2324-9250. DOI: 10.1029/2018eo095649.

Shen, Chaopeng, Xingyuan Chen, and Eric Laloy (2021). "Editorial: Broadening the Use of Machine Learning in Hydrology". In: *Frontiers in Water* 3. ISSN: 2624-9375. DOI: 10.3389/frwa.2021.681023.

Shen, Chaopeng, Eric Laloy, Amin Elshorbagy, Adrian Albert, Jerad Bales, Fi-John Chang, Sangram Ganguly, Kuo-Lin Hsu, Daniel Kifer, Zheng Fang, Kuai Fang, Dongfeng Li, Xiaodong Li, and Wen-Ping Tsai (2018). "HESS Opinions: Incubating Deep-Learning-Powered Hydrologic Science Advances as a Community". In: *Hydrol. Earth Syst. Sci.* 22.11, pp. 5639–5656. ISSN: 1607-7938. DOI: 10.5194/hess-22-5639-2018.

Shen, Chaopeng and Kathryn Lawson (2021). "Applications of Deep Learning in Hydrology". In: *Deep Learning for the Earth Sciences*. Ed. by Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein. 1st ed. Wiley, pp. 283–297. ISBN: 978-1-119-64614-3 978-1-119-64618-1. DOI: 10.1002/9781119646181.ch19.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis (2017). "Mastering the Game of Go without Human Knowledge". In: *Nature* 550.7676 (7676), pp. 354–359. ISSN: 1476-4687. DOI: 10/gcsmk9.

Sit, Muhammed, Bekir Z. Demiray, Zhongrun Xiang, Gregory J. Ewing, Yusuf Sermet, and Ibrahim Demir (2020). "A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources". In: *Water Science and Technology* 82.12, pp. 2635–2670. ISSN: 0273-1223. DOI: 10/ghwnnk.

SNO KARST (2021). *Time Series of Type Hydrology-Hydrogeology in Le Lez (Méditerranée) Basin - MEDYCYSS Observatory - KARST Observatory Network - OZCAR Critical Zone Network Research Infrastructure*. In collab. with Christelle Batiot-Guilhe, Jean-Luc Seidel, Jean-François Boyer, Pascal Brunet, Véronique De Montety, Frédéric Hernandez, Hervé Jourde, Hervé Jourde, Véronique

Léonardi, Pierre Marchand, Rémi Muller, Nicolas Patris, Aurore Remes-Busiau, Nathalie Rouché, Jean-Denis Taupin, Juliette Fabre, and Olivier Lobry. DOI: 10.15148/CFD01A5B-B7FD-41AA-8884-84DBDDAC767E.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958. ISSN: 1533-7928. URL: http://jmlr.org/papers/v15/srivastava14a.html (visited on 11/14/2021).

Stevanović, Zoran (2019). "Karst Waters in Potable Water Supply: A Global Scale Overview". In: *Environ Earth Sci* 78.23, p. 662. ISSN: 1866-6299. DOI: 10.1007/s12665-019-8670-9.

Stoll, S., H. J. Hendricks Franssen, R. Barthel, and W. Kinzelbach (2011). "What Can We Learn from Long-Term Groundwater Data to Improve Climate Change Impact Studies?" In: *Hydrol. Earth Syst. Sci.* 15.12, pp. 3861–3875. ISSN: 1607-7938. DOI: 10.5194/hess-15-3861-2011.

Sudheer, K. P., P. C. Nayak, and K. S. Ramasastri (2003). "Improving Peak Flow Estimates in Artificial Neural Network River Flow Models". In: *Hydrological Processes* 17.3, pp. 677–686. ISSN: 1099-1085. DOI: 10.1002/hyp.5103.

Supreetha, B. S., Narayan Shenoy, and Prabhakar Nayak (2020). "Lion Algorithm-Optimized Long Short-Term Memory Network for Groundwater Level Forecasting in Udupi District, India". In: *Applied Computational Intelligence and Soft Computing* 2020, pp. 1–8. ISSN: 1687-9724, 1687-9732. DOI: 10.1155/2020/8685724.

Sutton, Richard S. and Andrew G. Barto (1998). *Reinforcement Learning: An Introduction*. Red. by Francis Bach. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: A Bradford Book. 344 pp. ISBN: 978-0-262-19398-6.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus (2014). *Intriguing Properties of Neural Networks*. arXiv: 1312.6199 [cs]. URL: http://arxiv.org/abs/1312.6199 (visited on 01/26/2022).

Taylor, Richard G., Bridget Scanlon, Petra Döll, Matt Rodell, Rens van Beek, Yoshihide Wada, Laurent Longuevergne, Marc Leblanc, James S. Famiglietti, Mike Edmunds, Leonard Konikow, Timothy R. Green, Jianyao Chen, Makoto Taniguchi, Marc F. P. Bierkens, Alan MacDonald, Ying Fan, Reed M. Maxwell, Yossi Yechieli, Jason J. Gurdak, Diana M. Allen, Mohammad Shamsudduha, Kevin Hiscock, Pat J.-F. Yeh, Ian Holman, and Holger Treidel (2012). "Ground Water and Climate Change". In: *Nature Climate Change* 3, p. 322. URL: https://doi.org/10.1038/nclimate1744.

Tebaldi, Claudia, Kevin Debeire, Veronika Eyring, Erich Fischer, John Fyfe, Pierre Friedlingstein, Reto Knutti, Jason Lowe, Brian O'Neill, Benjamin Sanderson, Detlef van Vuuren, Keywan Riahi, Malte Meinshausen, Zebedee Nicholls, Katarzyna B. Tokarska, George Hurtt, Elmar Kriegler, Jean-Francois Lamarque, Gerald Meehl, Richard Moss, Susanne E. Bauer, Olivier Boucher, Victor Brovkin, Young-Hwa Byun, Martin Dix, Silvio Gualdi, Huan Guo, Jasmin G. John, Slava Kharin, YoungHo Kim, Tsuyoshi Koshiro, Libin Ma, Dirk Olivié, Swapna Panickal, Fangli Qiao, Xinyao Rong, Nan Rosenbloom, Martin Schupfner, Roland Séférian, Alistair Sellar, Tido Semmler, Xiaoying Shi, Zhenya Song, Christian Steger, Ronald Stouffer, Neil Swart, Kaoru Tachiiri, Qi Tang, Hiroaki Tatebe, Aurore Voldoire, Evgeny Volodin, Klaus Wyser, Xiaoge Xin, Shuting Yang, Yongqiang Yu, and Tilo Ziehn (2021). "Climate Model Projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6". In: *Earth Syst. Dynam.* 12.1, pp. 253–293. ISSN: 2190-4987. DOI: 10.5194/esd-12-253-2021.

Thierion, Charlotte, Laurent Longuevergne, Florence Habets, Emmanuel Ledoux, Philippe Ackerer, Samer Majdalani, Etienne Leblois, Simon Lecluse, Eric Martin, Solen Queguiner, and Pascal Viennot (2012). "Assessing the Water Balance of the Upper Rhine Graben Hydrosystem". In: *Journal of Hydrology* 424–425, pp. 68–83. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2011.12.028.

Thiéry, D and P Bérard (1983). *Alimentation en eau de la ville de Montpellier - captage de la source du Lez - études des relations entre la source et son réservoir aquifère*. BRGM No. 83, SNG 167 LRO. URL: http://infoterre.brgm.fr/rapports/83-SGN-167-LRO.pdf (visited on 06/30/2021).

Thober, Stephan, Andreas Marx, and Friedrich Boeing (2018). *Auswirkungen der globalen Erwärmung auf hydrologische und agrarische Dürren und Hochwasser in Deutschland*, p. 20.

Toth, E (2009). "Classification of Hydro-Meteorological Conditions and Multiple Artificial Neural Networks for Streamflow Forecasting". In: *Hydrol. Earth Syst. Sci.*, p. 12. DOI: 10.5194/hess-13-1555-2009.

Toth, E. (2013). "Catchment Classification Based on Characterisation of Streamflow and Precipitation Time Series". In: *Hydrology and Earth System Sciences* 17.3, pp. 1149–1159. ISSN: 1607-7938. DOI: 10.5194/hess-17-1149-2013.

Tsai, Wen-Ping, Kuai Fang, Xinye Ji, Kathryn Lawson, and Chaopeng Shen (2020). "Revealing Causal Controls of Storage-Streamflow Relationships With a Data-Centric Bayesian Framework Combining Machine Learning and Process-Based Modeling". In: *Frontiers in Water* 2. ISSN: 2624-9375. DOI: 10.3389/frwa.2020.583000.

UBA (2020). *Trockenheit in Deutschland – Fragen und Antworten*. Umweltbundesamt. URL: https://www.umweltbundesamt.de/themen/trockenheit-in-deutschland-fragen-antworten (visited on 01/25/2021).

UFZ (2021). *UFZ Dürremonitor Deutschland*. URL: https://www.ufz.de/index.php?de=37937.

UN-ECOSOC (2021). *Progress towards the Sustainable Development Goals*. E/2021/58. URL: https://undocs.org/en/E/2021/58 (visited on 01/25/2022).

UN-Water, ed. (2020). *Water and Climate Change*. The United Nations World Water Development Report 2020. Paris: UNESCO. 219 pp. ISBN: 978-92-3-100371-4.

UNESCO (2012). *World's Groundwater Resources Are Suffering from Poor Governance*. URL: http://www.unesco.org/new/en/media-services/single-view/news/worlds_groundwater_resources_are_suffering_from_poor_gove/ (visited on 05/11/2020).

UNFCCC (2021). *Nationally Determined Contributions under the Paris Agreement; Synthesis Report by the Secretariat*. FCCC/PA/CMA/2021/8. URL: https://unfccc.int/documents/306848.

Van Rossum, Guido (1995). *Python Tutorial*.

Van, Song Pham, Hoang Minh Le, Dat Vi Thanh, Thanh Duc Dang, Ho Huu Loc, and Duong Tran Anh (2020). "Deep Learning Convolutional Neural Network in Rainfall–Runoff Modelling". In: *Journal of Hydroinformatics* 22.3, pp. 541–561. ISSN: 1464-7141. DOI: 10.2166/hydro.2020.095.

Van Wynsberghe, Aimee (2021). "Sustainable AI: AI for Sustainability and the Sustainability of AI". In: *AI Ethics* 1.3, pp. 213–218. ISSN: 2730-5961. DOI: 10.1007/s43681-021-00043-6.

Van der Walt, Stéfan, S Chris Colbert, and Gaël Varoquaux (2011). "The NumPy Array: A Structure for Efficient Numerical Computation". In: *Comput. Sci. Eng.* 13.2, pp. 22–30. ISSN: 1521-9615. DOI: 10.1109/mcse.2011.37.

Vesanto, Juha (2005). *SOM Toolbox: Implementation of the Algorithm*. SOM toolbox 2.0 implementation. URL: http://www.cis.hut.fi/projects/somtoolbox/documentation/somalg.shtml (visited on 04/04/2018).

Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini (2020). "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals". In: *Nat Commun* 11.1 (1), p. 233. ISSN: 2041-1723. DOI: 10.1038/s41467-019-14108-y.

Vogt, Frédéric P.A. (2021). *Mannkendall/Python*. Version v1.1.0. Zenodo. DOI: 10.5281/ZENODO.4495590.

Wang, Xiaozhe, Kate A. Smith, and Rob J. Hyndman (2006). "Characteristic-Based Clustering for Time Series Data". In: *Data Mining and Knowledge Discovery* 13.3, pp. 335–364. ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-005-0039-x.

Wegehenkel, Martin and Kurt-Christian Kersebaum (2009). "An Assessment of the Impact of Climate Change on Evapotranspiration, Groundwater Recharge, and Low-Flow Conditions in a Mesoscale Catchment in Northeast Germany". In: *Journal of Plant Nutrition and Soil Science* 172.6, pp. 737–744. ISSN: 1522-2624. DOI: 10.1002/jpln.200800271.

Wriedt, Gunter (2020). *Grundwasserbericht Niedersachsen: Sonderausgabe Zur Grundwasserstandssituation in Den Trockenjahren 2018 Und 2019*. 41. NLWKN.

Wu, Wen-Ying, Min-Hui Lo, Yoshihide Wada, James S. Famiglietti, John T. Reager, Pat J.-F. Yeh, Agnès Ducharne, and Zong-Liang Yang (2020). "Divergent Effects of Climate Change on Future

Groundwater Availability in Key Mid-Latitude Aquifers". In: *Nat Commun* 11.1, p. 3710. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17581-y.

Wunsch, Andreas (2021). *Supporting Information*. URL: doi.org/10.5281/zenodo.5645467 (visited on 11/05/2021).

Wunsch, Andreas and Tanja Liesch (2020). *Entwicklung und Anwendung von Algorithmen zur Berechnung von Grundwasserständen an Referenzmessstellen auf Basis der Methode Künstlicher Neuronaler Netze*. Fachbericht. Karlsruhe Institute of Technology, p. 191. DOI: 10.5445/IR/1000136522.

Wunsch, Andreas, Tanja Liesch, and Stefan Broda (2018). "Forecasting Groundwater Levels Using Nonlinear Autoregressive Networks with Exogenous Input (NARX)". In: *Journal of Hydrology* 567, pp. 743–758. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2018.01.045.

– (2021). "Groundwater Level Forecasting with Artificial Neural Networks: A Comparison of Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), and Non-Linear Autoregressive Networks with Exogenous Input (NARX)". In: *Hydrology and Earth System Sciences* 25.3, pp. 1671–1687. DOI: 10.5194/hess-25-1671-2021.

– (2022a). "Deep Learning Shows Declining Groundwater Levels in Germany until 2100 Due to Climate Change". In: *Nat Commun* 13.1 (1221). ISSN: 2041-1723. DOI: 10.1038/s41467-022-28770-2.

– (2022b). "Feature-Based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing Map-Ensembles". In: *Water Resour Manage* 36.1, pp. 39–54. ISSN: 0920-4741, 1573-1650. DOI: 10.1007/s11269-021-03006-y.

Wunsch, Andreas, Tanja Liesch, Guillaume Cinkus, Nataša Ravbar, Zhao Chen, Naomi Mazzilli, Hervé Jourde, and Nico Goldscheider (2022c). "Karst Spring Discharge Modeling Based on Deep Learning Using Spatially Distributed Input Data". In: *Hydrology and Earth System Sciences* 26.9, pp. 2405–2430. ISSN: 1027-5606. DOI: 10.5194/hess-26-2405-2022.

WWAP (2015). *Water for a Sustainable World*. The United Nations World Water Development Report 6.2015. United Nations World Water Assessment Programme - UNESCO. 122 pp. ISBN: 978-92-3-100071-3 978-92-3-100099-7. URL: http://www.unesco.org/new/en/loginarea/natural-sciences/environment/water/wwap/wwdr/2015-water-for-a-sustainable-world/ (visited on 05/11/2020).

Yin, Wenpeng, Katharina Kann, Mo Yu, and Hinrich Schütze (2017). *Comparative Study of CNN and RNN for Natural Language Processing*. arXiv: 1702.01923 [cs]. URL: http://arxiv.org/abs/1702.01923 (visited on 01/29/2022).

Zeiler, Matthew D. and Rob Fergus (2014). "Visualizing and Understanding Convolutional Networks". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, pp. 818–833. ISBN: 978-3-319-10590-1.

Zhang, Andi, James Winterle, and Changbing Yang (2020). "Performance Comparison of Physical Process-Based and Data-Driven Models: A Case Study on the Edwards Aquifer, USA". In: *Hydrogeol J*. ISSN: 1431-2174, 1435-0157. DOI: 10.1007/s10040-020-02169-z.

Zhang, Jianfeng, Yan Zhu, Xiaoping Zhang, Ming Ye, and Jinzhong Yang (2018). "Developing a Long Short-Term Memory (LSTM) Based Model for Predicting Water Table Depth in Agricultural Areas". In: *Journal of Hydrology* 561, pp. 918–929. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2018.04.065.

Zhang, Juan, Xiaoying Zhang, Jie Niu, Bill X. Hu, Mohamad Reza Soltanian, Han Qiu, and Lei Yang (2019). "Prediction of Groundwater Level in Seashore Reclaimed Land Using Wavelet and Artificial Neural Network-Based Hybrid Model". In: *Journal of Hydrology* 577, p. 123948. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2019.123948.

Zhao, Ling, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li (2020). "T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction". In: *IEEE Transactions on Intelligent Transportation Systems* 21.9, pp. 3848–3858. ISSN: 1558-0016. DOI: 10.1109/TITS.2019.2935152.

Zhu, Jiawei, Qiongjie Wang, Chao Tao, Hanhan Deng, Ling Zhao, and Haifeng Li (2021). "AST-GCN: Attribute-Augmented Spatiotemporal Graph Convolutional Network for Traffic Forecasting". In: *IEEE Access* 9, pp. 35973–35983. ISSN: 2169-3536. DOI: 10/gm8pbn.